CLASSIFICATION OF MULTI-SOURCE SATELLITE IMAGES FOR LARGE SCALE LAND COVER MAPPING

Enric Juan

A Master Thesis in
The Centre d'Etudes Spatiales de la BIOsphère

Presented in partial fullfilment of the requirements
for the Master's Degree in Telecommunications
Engineering at Escola Tècnica d'Enginyeria de
Telecomunicació de Barcelona

Supervisors: Silvia Valero, Philippe Salembier

Jan 2018

Abstract

Satellite remote sensing imagery represents an attractive data source to monitor large regions with frequent updates. In this context, the operational production of accurate land cover maps plays an important role in global-scale environmental assessments and becomes crucial for a wide range of research domains. New earth observation missions such as Sentinel provide images with high spatial and temporal resolution. Accordingly, new image classification methods for the generation of reliable land cover maps are needed.

In the framework of the Sentinels Synergy for Agriculture (SENSAGRI) project at *Centre d'Études Spatial de la Biosphere* (CESBIO) in Toulouse (France), this work aims to describe new schemes for detecting crop areas along the agricultural season. The research has focused on performing statistical fusion at decision-level to combine classification results in order to exploit the synergies between Sentinel-1 and Sentinel-2 image times series.

To my parents, because I would never have come so far without them. And to Melisa, because the adventure would seem that comes to an end but it has just started. Because no one has taught me more than them.

Acknowledgements

I would like to express my very great appreciation to Silvia because the world needs lecturers like her.

I am also grateful to Philippe for his dedication and helpful advice.

I would also like to thank my colleagues from my internship at CESBIO, Milena, Ludovic and Eric, for their wonderful collaboration. But specially to the interns team for their support throughout this period.

Finally, I would also like to thank to Roberto and Anaís, because Toulouse will be always special thanks to them.

Revision history and approval record

Revision	Date	Purpose
0	02/02/2018	Document creation
1	31/03/2018	Document revision
2	15/04/2018	Document revision
3	26/04/2018	Final version

DOCUMENT DISTRIBUTION LIST

Name	e-mail
Enric Juan	enric.juan.92@gmail.com
Silvia Valero	silvia.valero@cesbio.cnes.fr
Philippe Salembier	philippe.salembier@upc.edu

Written by:		Reviewed and approved by:	
Date	26/04/2018	Date	26/04/2018
Name	Enric Juan	Name	Philippe Salembier
Position	Project Author	Position	Project Supervisor

Contents

List of Figures

List of Tables

Part I

Introduction

Chapter 1

Context

For several decades, Earth Observation (EO) makes possible a better understand of our planet. Then in the global change era, the characterization of the dynamics related to the transformation of continental surfaces - consumption of agricultural areas, deforestation or urban sprawl - is essential.

In this context, space remote sensing offers the possibility of frequently mapping the entire planet. More specifically, images from satellite acquisitions produce maps that give a graphical representation of land surfaces such as land cover-land use.

1.1 Land cover and use maps

Land cover refers to the physical and biological cover over the surface of land, including water, vegetation, bare soil, and/or artificial structures. Land use denotes how humans use the biophysical or ecological properties of land. Land use is characterized by the arrangements, activities and inputs people undertake in a certain land cover type to produce, change or maintain it. Definition of land use in this way establishes a direct link between land cover and the actions of people in their environment.

Land cover mapping has been recognized as a fundamental task for deriving information for scientific, environmental management and policy purposes at global, regional and local scales [1, 2]. Information on the characteristics and use of land surface elements has proved crucial for environmental studies involving bio-geo-chemical cycles, conservation and the management of natural resources, urban planning, food and health among others [3, 4, 5].

Remote sensing (RS) is the most significant technology for effective land cover mapping at large scales, bringing numerous advantages such as cost-effectiveness and repeatability of observations [6]. For instance, Landsat images have served a great deal in the classification of different landscape components at a larger scale [7]. Also, mapping activities have widely exploited optical [8] and radar [9, 10] satellite

imagery for classification and the mapping of land cover–land use (LCLU) .

1.2 Remote sensing and satellite images

Radar and optical remote sensing data deliver complementary information, hence land cover classification tasks can take advantage of it. For instance, optical energy reflected by vegetation is dependent on leaf structure, pigmentation and moisture. Optical products are commonly available as multi-spectral images consisting of several bands of data, which can offer different information on land properties based on its spectral reflectance. In contrast, active microwave energy scattered by vegetation is dependent on the size, density, orientation and dielectric properties of elements comparable to the size of the radar wavelength. Radar signals are typically only generated at a single wavelength for each sensor, and interact in a characteristic way with structural land properties. Multiple bands can be composed with the polarized combinations of the scattered signals (*e.g.*, HH plus VV). Furthermore, techniques such as interferometric SAR (InSAR), make use of differential phases of reflected signals to detect land surface changes.

[11, 12] found that SAR-based texture information combined with Visible and Near-Infrared (VNIR) optical data can improve the classification of vegetation areas. ALOS PALSAR[1] and phenological information from the MODIS [2] optical sensor were used by [13] to map forest areas using decision tree algorithms. Optical sensors may be disturbed by cloud presence along the line of sight. In order to avoid this effect L-band SAR data were used and improved separability of evergreen shrubs and crops from forests. Therefore, the use of radar and optical satellite data for classification tasks may imply a significant advantage.

The Sentinels missions

Nowadays, new opportunities arise for Earth Observation thanks to new European satellite missions such as the Sentinel-1, -2 satellites. This ESA mission provides high operational capability, long-term continuity, superior calibration of sensors and a variety of sensing methods and products for the scientific community [14]. Also, Sentinel data distribution is supported by the key advantage of a full free and open access policy for the majority of the products.

On the one hand, the Sentinel-1 sensors provide C-band SAR images in single/dual polarization for a variety of acquisition modes [15]. Its combination of high spatial (5 $\times$ 20 m in the Interferometric Wide Swath mode), large coverage (up to 400 km) and improved temporal resolution is providing accurate land cover mapping. Frequent revisit time is a major advantage over previous radar missions, especially for the mapping and analysis of phenological dynamics in vegetation and agricultural land

[1]Phased Array type L-band Synthetic Aperture Radar
[2]NASA Imaging satellite mission

covers, together with the dual polarization capability and rapid product delivery [15].

On the other hand, the Sentinel-2 Multi-spectral Instrument (MSI) optical sensors provide radiometrically and geometrically superior multi-spectral high spatial resolution images over the global surface, at high revisit time (5 days at the Equator with two satellites in orbit) and a wide field of view covering 290 km with 13 bands in the optical Near-Infrared (NIR) , Short-Wavelength Infrared (SWIR) parts of the electromagnetic spectrum [16]. Major advantages of Sentinel-2 with respect to previous satellite missions (*e.g.* Landsat series) are given by higher spatial and spectral content. Besides, for vegetation mapping Sentinel-2 is especially relevant for the presence of two new bands in the red edge spectrum, at 705 and 740 nm [17].

1.3 Statement of the problem

In the context of land cover mapping, decametric or metric spatial resolution imagery is needed in order to produce detailed maps. Besides, many land cover classes can only be recognized by their temporal dynamics and therefore, high temporal resolution is also needed.

As presented in the previous section, optical as well as radar remote sensing technologies are currently undergoing a very significant evolution in terms of the quality and the quantity of information. In particular, sensors are becoming more precise and can capture data at a very high resolution. This can be observed in terms of spatial and spectral resolutions. Besides, in order to understand how Earth is changing a high revisit time is also necessary. In this context, the use of Satellite Image Time Series (SITS) allows to obtain large data series with short time intervals between images taken from the same scene.

Therefore, high temporal components integrated with spectral and spatial dimensions allows the identification of complex patterns concerning analysis of land-cover dynamics. This wealth of information is very interesting from the application viewpoint but it also generates real challenges.

1.3.1 Challenges

The availability of Sentinels imagery [16], with its unique characteristics, enable the implementation of accurate land cover maps. This production systems involves the delivery of up to date and accurate information [18]. Therefore, the high spatial and temporal resolution of these new time series of satellite images is an asset for the characterization of land occupations that evolve over time. But, those new opportunities involves several challenges since the processing of these time series images involve the management of large volumes of data never studied before.

The state of the art in land cover mapping uses image classification [19]. Super-

vised classification is known to be superior to unsupervised approaches. Its main drawback is the need of training data (ground truth or other reference data). This has often been criticized and seen as a barrier for the implementation of global scale operational systems. However, it has been shown that some supervised classifiers are robust to errors in the training data and therefore could use slightly out-of-date reference data for the training step [20].

In this context, several classification methods have been successfully applied for land use mapping [21]. However these methods have rarely used all the information provided by the new time series of satellite images since the classification algorithms may suffer from the high dimensionality of the input data [3].

For many applications of image classification problems, the information provided by a single sensor may be incomplete resulting in misclassification. Besides, the algorithm that is effective for one data set may be unsuitable to other data sets. Hence, multiple classifier combination may outperform any individual classifier by integrating the advantages of various classifiers.

The development of adequate fusion techniques is an important ongoing field of research. In general, the fusion aim is to generate information of "greater quality". Images acquired over the same site by different sensors are partially redundant, as they represent the same scene. Also they are partially complementary, since the sensors have different characteristics and physical interaction mechanisms are different. Therefore, fusion techniques in land cover classification can help reducing the imprecision and it can provide a more complete description resulting in a better classification.

1.3.2 Objectives

The main goal of this work is to improve the accuracy of produced land cover maps resulting by the classification of remote sensing data. More specifically, this work is focused on the improvement of existing classification processing chain by proposing the integration of the new Sentinel-1 and -2 satellites image time series. To perform it, several fusion strategies have been proposed in this work to merge the single radar and optical classifications. The purpose of such strategies have been to exploit the use of the probabilistic output enabled by the Random Forest algorithm.

Under this purpose, the first work here is a study based on the analysis of the predicted results from single radar and optical classifiers. This study is mainly performed from a statistical viewpoint.

Therefore, the second study is the presentation of the fusion strategies. The interest of this work is to show a experimental evaluation of the fusion strategies. Besides, the results have been compared with other classification strategies in other to present the advantages of the fusion framework.

[3]The dimension of a classification problem is given by the number of features that compose the input samples.

Finally, the third study presented in this work is based on the improvement of the Random Forest performance. In this context, the purpose of this work is to obtain a more accurate probability vector by means of a weighting step in the Random Forest algorithm.

1.3.3 Report structure

This work is structured as follows:

Part I. It details the context of this research work. The land cover map definition is given. Also, the construction of these maps by using supervised classification approaches is presented. Besides, the challenges involving the classification of new time series satellite images and the integration of remote sensing data is described. Finally, the main goals of this work are detailed.

Part II. It presents the methodology and data which have been used. A description of the input data used in this work is given in Chapter 2. The supervised classification of satellite images is detailed in Chapter 3. Then, Chapter 4 presents the classification chain for the single classifiers. Finally, a theoretical definition of the proposed fusion methods is given by Chapter 5.

Part III. It consists in the description of the experimental results obtained throughout this project. Chapter 6 presents the statistical analysis provided by the new features implemented for the classification algorithm. An evaluation of the different fusion approaches is given in Chapter 7 introducing a set of metrics and visualization tools for the study. Finally, in Chapter 8 a new strategy for the classification chain is given presenting the evaluation results.

Part IV. It mainly contains Chapter 9 presenting the general conclusion. This chapter summarizes the main results of this project, and highlights the most important conclusions. Besides, a set of future perspectives is discussed.

Part II

Methodology and data

Chapter 2

Description of the input Data

As commented in Chapter 1, the main purpose of this work is the integration of the complementary information obtained from single classifiers. The goal of each single supervised classifier is to predict the pixel labels to produce land cover maps. The input data used in each single classification task is different. In one case, the Sentinel-2 optical image times series whereas in the second case the input data corresponds to Sentinel-1 radar data. Both data cover the same location area around the south of France.

The details of the reference data used to train the supervised classifiers are also detailed in a second section. In it, the description of the training and validation data extracted from the global reference data set is described.

This chapter contains three different sections. The studied area for classification is presented in the first section. This large area is composed by three satellite images times series. The details of the reference data used to train the supervised classifiers are also detailed in a second section. In it, the description of the training and validation data extracted from the global reference data set is described. The reference data is composed by the classes of interest for classification purposes, these classes are presented and the origin of this data is detailed. In the last section, the Sentinel's satellite data is presented. The characteristics of the images are described by both data sources. Besides, the pre-processing tasks that have been carried out on the raw data are also presented.

2.1 Study area

In the following section, a description of the study areas is exposed. Three different images (also known as tiles T30TYP, T31TCJ and T31TCH) provided by the Sentinel-1 and the Sentinel-2 satellites has been used for this work. The three tiles are located on around Toulouse, such as it can be seen on Figure 2.2.

Images time series cover an agricultural year, from October 2015 to October 2016.

In the case of radar image times series, the number of acquired images at this agricultural year is equal to 84 dates, from October the 8th 2015 to December the 31st 2016. In contrast, the optical image times series only contains 33 acquisition dates, from November the 29th 2015 to October the 14st 2016, since Sentinel-2 has a a lower temporal resolution than Sentinel-1. Figure 2.1 shows the distributions of the radar and optical acquisition dates along the year, whereas each date implies one satellite image.

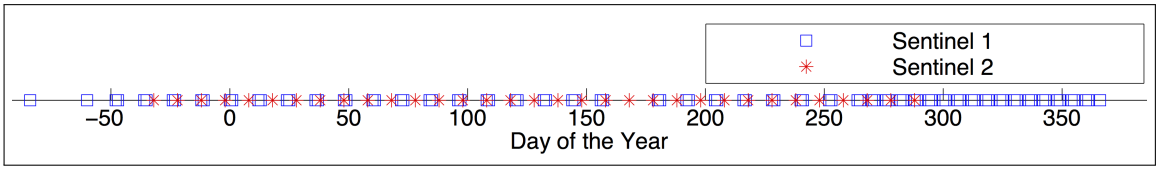


Figure 2.1: Distribution of the radar and optical acquisitions during the agricultural season 2015-2016.

It should be noticed that the optical acquisition dates represented on Figure 2.1 correspond to the T31TCH and T31TCJ tiles located on the right of the Figure. As Figure 2.2 shows, T30TYP data has been captured in a different orbit, therefore, the acquisition dates for this tile were different. In order to work with a regular temporal grid for the three zones, a linear interpolation has been applied on the T30TYP satellite data. Finally, the re-projection of radar and optical data were also performed to avoid pixel discontinuities between the tiles. The preprocessing is further discussed in Section 2.3.

Figure 2.2: The studied area composed by three zone-tiles on the south of France.

2.2 Reference data

Reference data is needed to train the classification system and the subsequent validation of the land cover/use maps. Such data must completely represent the true

landscape by capturing the diversity of crop types and land cover classes. Note that the classifier is not capable of identifying what is unknown in the training data set. The quality and the number of samples per map class are, both, very crucial factors for the accuracy assessment of the resulting crop maps. The reference data is composed of crop type and a significant set of "non-cropland" polygons (i.e. other land cover classes). Table 2.1 presents the culture classes divided by *Cropland* and *Non-cropland* [1].

The crop type must be identified by field observation during the corresponding growing season for each sampled parcel, field or piece of land larger than 0.25 ha with a minimum width of 30 m. Ideally, such information is provided several times along the agricultural season (intermediate-cover crops / bicultural cropping systems).

The reference data used to label the training and validation sets of pixels come essentially from French National databases such as *Institut Géographique National* (IGN) and from in-situ measurements made by field experts from the CESBIO. For tests of statistical confidence, 10 sets of pixel (*nbruns*) where constructed at random in order to match with the proceedings explained in Section 3.3. Figure 2.3 presents a visual example of the so-called reference data. As it could be noticed, it is composed by different polygons divided by colors which corresponds to the different culture classes. The use and treatment of the reference data is addressed further in Chapter 4.

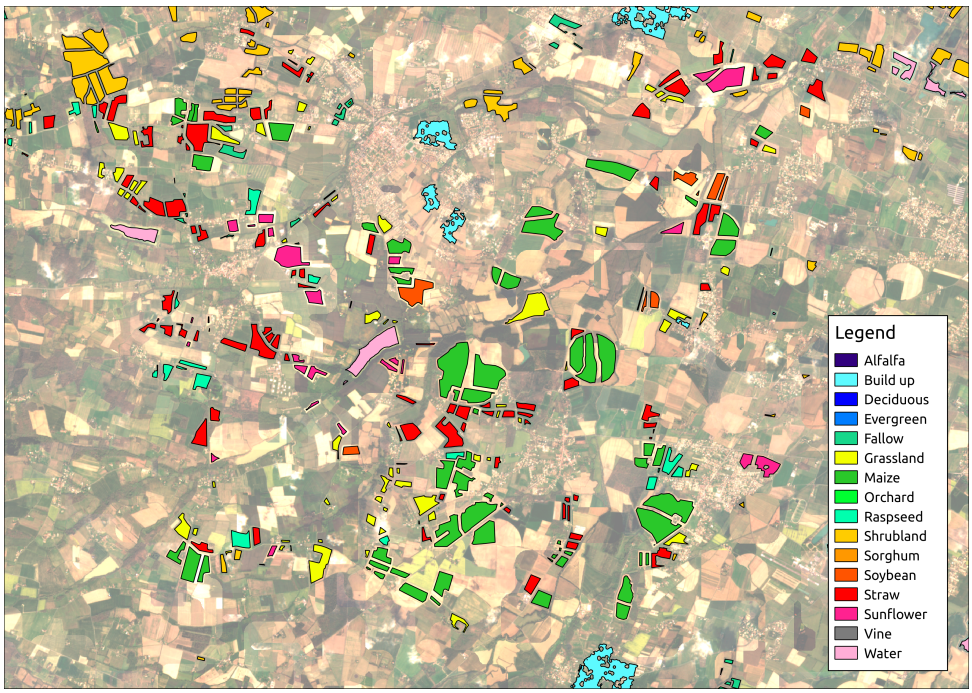


Figure 2.3: Reference data polygons set example

A summary of the set of pixels that composed the reference data is given in Table 2.1, presenting the culture classes and the pixels distributions among them.

[1]Because of its confusion and similarity with the class *Grassland*, it was decided to consider *Alfalfa* as a no-crop when it was constructed the three-tile zone dataset.

The training sets were constructed containing exactly 2000 pixels per classe. The validation sets were created too. When it is not possible to reach the good number because of a lack of representativity of some classes, the smallest number among the 10 random draws is taken.

Land cover	Cropland	Training	All Pixels	Distribution (%)
Vine	No	2000	7×10^5	1
Straw cereals	Yes	2000	5×10^5	10
Maize	Yes	2000	4×10^5	4
Sorghum	Yes	2000	3×10^4	0.3
Soybean	Yes	2000	5×10^4	0.3
Sunflower	Yes	2000	2×10^5	5.0
Alfalfa	Yes	2000	3×10^5	1.0
Grassland	No	2000	9×10^5	29.5
Fallow	No	2000	2×10^5	1.6
Shrubland	No	2000	1×10^5	3.9
Rapeseed	Yes	2000	8×10^4	1.0
Deciduous	No	2000	2×10^5	26.0
Evergreen	No	2000	2×10^4	5.0
Build up	No	2000	3×10^5	7.0
Water	No	2000	1×10^5	4.0
Orchard	No	2000	4×10^4	0.4

Table 2.1: Input pixels from the three tiles set: the first column indicates the number of pixels used for the learning step. The second column indicates the number of pixels per class available for the validation. The third column indicated the proportion of pixels per classes in the tested area.

2.3 Satellite data

The proposed work relies on a supervised classification system which exploits the different multi-temporal Sentinel missions: the high spatial-spectral resolution of Sentinel-2 and the weather independent acquisitions of Sentinel-1.

This section presents a brief description of the Sentinel-1 and -2, their RS instruments and the image preprocessing tasks to obtain the proper images for the classification chain.

2.3.1 Sentinel-1 images

As explained in Section 1.2, Sentinel-1 carries a single C-band synthetic aperture radar instrument which supports operation in dual polarization (HH+HV, VV+VH) in C band useful for land cover classification.

The input Sentinel-1 images are Level-1 Ground Range Detected (GRD) in Interferometric wide swath (IW). It consists of focused SAR data that has been detected, multilooked (5x1) and projected to ground range using an Earth ellipsoid model (WGS84). Pixel values represent the detected intensity only (phase information is not considered). The resulting product has approximately square resolution pixels square pixel spacing (10m x 10m at the mid-range value at mid-orbit altitude, averaged over all swaths) with a swath of 250 km. There is one image per polarisation channel. The incidence angle over the surveyed fields ranges between 35° to 43°.

Calibration

Sentinel-1 images are given in pixel Digital Numbers. These values need to be converted into the radar cross-section for distributed targets. This is done using the Look Up Table (LUT) in the metadata.

$$\sigma^0 = \frac{DN^2}{A_\sigma} \tag{2.1}$$

where A_σ is the radar cross-section LUT.

Orthorectification

Due to the tilt of the satellite sensor and the topographical variations of a scene, distances can be distorted in the SAR images (foreshortening, layover and shadow). Terrain corrections are needed to compensate for these distortions. Terrain correction is applied to geocode accurately the images using the digital elevation model from the Shuttle Radar Topography Missions at 30m for better accuracy [22].

Concatenation

Sentinel-1 images and Sentinel-2 images do not have the same swaths. To be able to use both of them in the classifier they need to cover the same area. Therefore, Sentinel-1 images are tiled into Sentinel-2 tiles, concatenating different Sentinel-1 images when necessary.

Filtering

A speckle filter [23] is applied to further reduce the speckle effect while preserving the 10 m spatial resolution and the fine structure present in the image. This filter produces images with reduced speckle effects from multi-temporal (more than 200 images) and multi-polarized (VH and VV).

2.3.2 Sentinel-2 images

Sentinel-2 carries a high-resolution optical instrument capable of sample ten bands in the VNIR and the SWIR , contributing to multi-spectral observations for applications such as land management or agriculture and forestry mapping among others.

The input Sentinel-2 images correspond to Orthorectified Top-Of-Atmosphere reflectances products. Then, top-of-atmosphere reflectances are converted to top-of-canopy reflectances by using MACCS-ATCOR Joint Algorithm [24] (named MAJA). This common software is a joint effort between CNES[2] teams working on MACCS (Multi-sensor Atmospheric Correction and Cloud Screening) software and DLR teams working ATCOR (Atmospheric and Topographic Correction) .

MACCS is a level 2A processor, which detects the clouds and their shadows, and estimates aerosol optical thickness (AOT) , water vapor and corrects for the atmospheric effects.

The use of the next Level 2A Sentinel-2 ten bands acquired in the VNIR and SWIR is proposed in Table 2.2.

Besides the Level 2A data, three masks identifying clouds, edges and saturation pixels are also used. These masks are grouped to construct an unique mask denoting the validity of the pixels.

[2] *Centre National d'Études Spatiales*

Spatial resolution (m)	Band number	Central Wavelength (nm)
10	2	490
	3	560
	4	665
	8	842
20	5	705
	6	740
	7	783
	8a	865
	11	1610
	12	2190

Table 2.2: Spatial and spectral Sentinel-2 resolutions

GapFilling: Linear interpolation

The purpose is to produce a reflectance image time series which is (i) gap-filled with respect to missing data (which can be due to clouds, cloud shadows and saturated pixels) and (ii) temporally sampled on a regular grid. The same approach proposed in [19], which consists of a linear interpolation of the invalid pixels using the previous and following cloud-free dates, is used here. The interpolation is applied over surface reflectance values and before the computation of derived spectral indices. The use of interpolation allows to estimate the values of the surface reflectances for any date, not only the dates of the invalid pixels. This allows the choice of a set of common dates for all the pixels of the area which in turn solves the issue of temporal shifts between adjacent satellite tracks.

Resampling

Looking at Table 2.2, it can be seen that the spectral bands of Sentinel-2 are captured at different spatial resolutions. In order to work with a unique spatial resolution, Sentinel-2 bands acquired at the spatial resolution of 20m are resampled to 10m.

Chapter 3

Supervised classification of satellite images

One of the main purposes of satellite remote sensing is to interpret the observed data and classify features. Then, the classification methods goal is to build a model able to predict for each pixel of the image a label which corresponds to a particular class. Coming from the field of machine learning, these algorithms are traditionally divided into two categories in the literature: supervised and unsupervised.

Unsupervised approaches (also known as clustering) seek to group similar samples within the same class, *e.g. k-Means*, *Self-Organizing Map* (SOM) or *Iterative Self-Organizing Data Analysis Technique yAy* (ISODATA) . Groups, also called clusters, are made up of similar samples that are dissimilar to samples belonging to other clusters. A class is then assigned *a posteriori* with each cluster. However, cluster recognition is a complex and time-consuming issue that can only be achieved by an expert from the study area. In many cases, post-processing is needed before labeling clusters, and matching them with the nomenclature [25]. In addition, pre-processing is also necessary to prevent classes with strong variance from dominating clusters (*e.g.* a seasonal crop). In addition, clustering methods are costly in time and computing resources as the size of the images increases. For all these reasons, supervised approaches are generally favored in the context of mapping over large areas [26].

This chapter focuses on the supervised classification framework. A first part introduces some generalities on supervised learning. A second part focuses on the use of supervised approaches for the classification of time series of satellite images. Then, a third part explains the main characteristics and the operating principle of the machine learning algorithm used during this work. Finally, it is described the evaluation process of the classification methods.

3.1 Introduction to the supervised learning

The goal of supervised learning is to automatically learn rules to predict the labels from a set of samples. The set of rules is learned from examples, or training samples, provided by reference data. Hence, by means of this learned rules, a classification model is constructed and it enables the creation of land use maps.

More precisely, Figure 3.1 details the entire classification process. The samples contained in the reference data are divided during the sampling step into two subsets. On the one hand, the learning (or training) samples are used as prior knowledge on land use. On the other hand, the test (or validation) samples are used in the evaluation phase, described in Section 3.4.

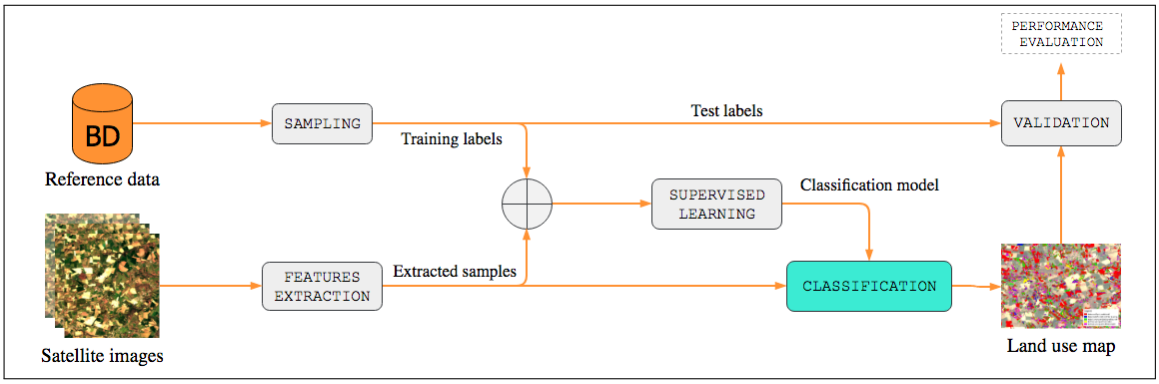


Figure 3.1: Supervised classification system

The central stage of the process is the supervised learning block. From the training labels and samples, the classification algorithm *learns* and establishes a prediction pattern called classification model. In other words, the decision rules defined by the model predict the land use classes for new samples. Ideally, the classification model should be able to generalize what it has learned to new samples. Hence, the learning process becomes critical in order to generate a suitable classification model.

One of the main difficulties of the learning phase is to find the compromise between a too simple model and a model that is too specific to the learning samples. The model may perfectly identify the learning samples, but be unable to correctly predict classes from new samples that were not used to build the model. Thus, when such case happens that a model learns the noise in the training data is called *Overfitting*. The problem is that these concepts do not apply to new data and negatively impact the model ability to generalize and, therefore in the model performance.

Oppositely, a too simple model is unable to capture the relevant relationships between the learning samples. This effect is known as *Underfitting*. In both cases, the constructed model is unable to generalize and predict the proper labels from the new samples.

This problem is also known as the bias-variance trade-off. The bias of an algorithm is characterized by its error on the set of learning data, whereas the variance of an

algorithm corresponds to the difference between the error made on the learning data and the error made on the test data. The bias-variance trade-off, therefore, consists in finding a balance between the complexity of the model and its capacity to generalize.

Learning samples play a vital role in achieving a good bias-variance trade-off. As shown in Figure 3.1, the learning samples are described by the values of the features vectors extracted from the satellite data. Classification algorithms then use these samples to build their decision rule. For these reasons, the learning samples must be representative of the population on which the model will be applied. They must describe the multiple appearances of the predicted classes.

Then, two important parameters of the problem of supervised classification are 1) n the number of learning samples, and 2) p the size of the feature vector. Increasing the values of n and p generally makes it possible to learn a more complex model while controlling the variance.

However, increasing the size of the vector p is not always a good solution. Some traditional methods, including statistical methods, are inefficient when the dimension of the problem p becomes too high, possibly greater than n. This phenomenon is known as the curse of dimensionality, or Hughes phenomenon, decreases the performance of classification algorithms [27]. In general, increasing the number of training samples provides a better description of class appearances and limits the curse of dimensionality. Thus, classification performance generally increases with the number of samples.

3.2 Supervised classification of time series

In the context of land use mapping, many methods have been proposed for the classification of satellite data. However, few studies focus on the classification of time series of optical satellite images. This lack of studies is due to the lack of quality reference data, and the recent availability of optical time series with high spatial resolution.

In this context, pattern classification algorithms are often categorized as parametric or non-parametric. Parametric methods require the knowledge of the statistics of the classification problem. If the probability of each class is known at any location in the d-dimensional pattern space, then an optimum classification of an unknown pattern can be made by selection of the most probable class at that point. But, the majority of parametric methods assume that the distribution of variables describing samples belonging to the same class follows a normal distribution, which is rarely the case in the context of time series classification. Thus, these approaches fail to take into account the different representations of certain classes, and the spectro-temporal variations present in the time series. For these reasons, non-parametric methods are more efficient than parametric methods when land-use class distributions are unknown [28]. Each supervised learning algorithm has its own advantages and disadvantages that could lead to different results on the same data.

And among the ensemble methods, Random Forest (RF) constructs a set of binary decision trees [29]. In the context of land use mapping, the number of works using RF is steadily increasing [30, 31, 32, 33]. In addition, the RF has many advantages: a reduced computing time due to the possibility of building trees in parallel [34], the ability to input large volumes of data (*i.e.* a large number of features) [35], and the possibility to visualize the built trees. But, each classification algorithm has its own advantages and disadvantages, and their performance depends on the datasets studied [36].

3.3 The Random Forest algorithm

The interest of using the RF algorithm, to classify optical times series has been recently corroborated [37]. Criteria such as accuracy, computational burden, processing time, stability or robustness were taken into account in order to select the more suitable algorithm.

The supervised classification algorithm chosen for the purpose of this project is the wide-known Random Forest [29]. The RF is a supervised learning algorithm based on the technique of the binary decision tree. The particularity of the RF is to combine a set of binary decision trees to build its decision rule. Thus, this section is focused in the description of the classification method used.

In a first part, the principle of ensemble methods is described. In a second part, the induction of the binary decision trees is explained. Finally, a third part is dedicated to the principle of the RF operation.

3.3.1 Ensemble methods

The idea of ensemble methods is to combine the predictions of different classification algorithms in order to obtain a more efficient classifier [38].

Two strategies exist for the construction of these methods: 1) to combine different classification algorithms, and 2) to combine different variants of the same classification algorithm. Then, the outputs of each classifier are merged, usually by a majority voting process [39]. Therefore, the interest of the ensemble methods relies in the use of classifiers who have considerably different behaviors. Empirically, the ensemble of classifiers tend to yield better results when there is a significant diversity among the models [40, 41, 42].

Then, three techniques exist for constructing the so-called ensemble methods from the same classification algorithm.

The first technique is the *random subspace* [43, 44]. In this approach, diversity is added using a subset of variables randomly drawn. In contrast, all samples are used to learn all classifiers. Thus, each classifier is specialized for a group of specific

variables. The combination of classifiers then makes it possible to obtain a reliable algorithm over the entire set of subspaces.

The other two techniques are the *bagging* and the *boosting*. For both cases, diversity is added by playing on the use of learning samples to build each classifier.

More specifically, the *bagging* (from *aggregating bootstrap*) is based on building several classifiers by means of subsets of different learning samples [45]. Each subset of learning samples, referred to as *bootstrap* samples, is obtained by using a randomly drawn subset of N samples from the learning set [46]. Combining the predictions from the built set of classifiers using the different *bootstrap* samples improves the ability to generalize for those classifiers experiencing *Overfitting*. This is due to the decrease of the variance of each classifier. Moreover, its building is easily parallelized.

Regarding the *boosting* techinque, it is based on the *weak classifiers* building. In the case of a binary classification, a classifier is said to be *weak* if it is wrong less than 50% of the chances. In this case, the set of classifiers is built recursively: each classifier is an adaptive version of the previous one where the wrong predicted samples are over-weighted. Thus, unlike *bagging*, classifiers may not be parallelized since are dependent on each other. The best-known algorithm based on this principle is *AdaBoost* (*Adaptive Boosting*) [47].

3.3.2 Binary decision trees

In the context of classification, decision trees are used to summarize a set of rules in a hierarchical tree structure. Their main advantage is to provide a graphical and intuitive representation of the decision rule that will determine the label of the predicted samples.

Figure 3.2 shows an example of a binary decision tree for the classification of five crop classes according to different information characterizing their phenological cycle. In the initial state, the tree is made up from the root that tests the date of the beginning of growth.

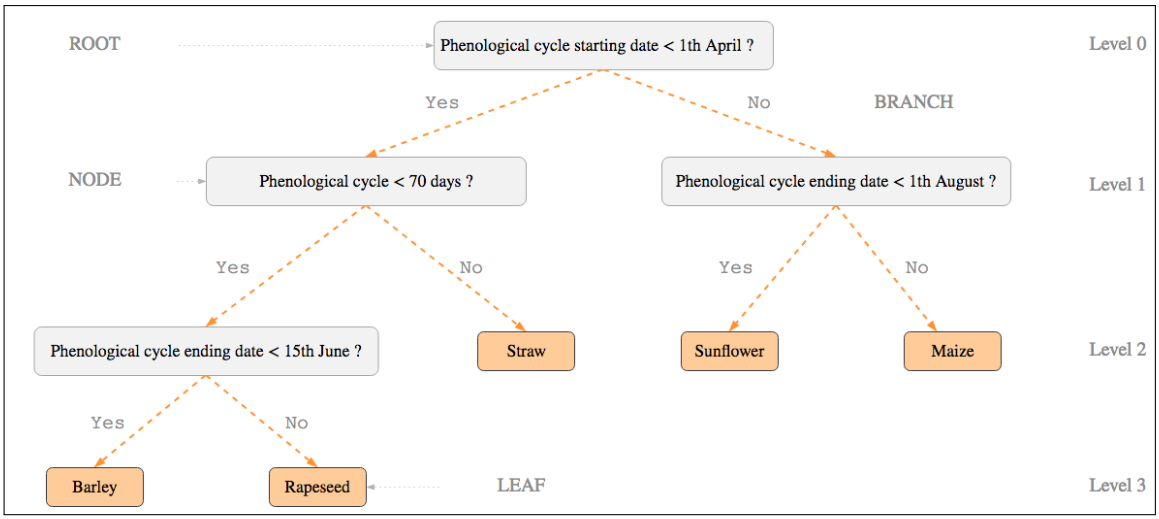


Figure 3.2: Example of a binary decision tree for the classification of crops according to different information characterizing their phenological cycle

If this last event took place before 1st April, then the sample takes the left branch. Otherwise, the sample follows the right branch. Each node is thus defined by the joint choice that will induce a partition into two subsets. Terminal nodes in orange, also called leaves, are nodes that do not have child nodes. They contain the final ranking decision. The term of level corresponds to the depth of the different nodes. By default, the root is at level 0.

The construction of a binary decision tree begins with the creation of the root that contains all the learning samples. The goal is to add new nodes that divide the samples into more homogeneous subsets. Ideally, a set of samples is homogeneous if the samples have similar behaviors (*i.e.* they belong to the same class for a classification problem). As shown in Figure 3.2, the construction continues on several levels until terminal nodes are obtained. Then, building a binary decision tree requires:

1. The definition of a splitting rule that makes it possible to divide the samples into more homogeneous subsets.
2. A rule to decide that a node is terminal.
3. A rule allowing the assignment of each leave to one of the classes. Generally, the class assigned to a leaf corresponds to the dominant class among the learning samples belonging to it.

The partitioning rule is associated with each node to distribute the samples in two child nodes. This rule is determined by means of an incoming feature and an attribute value test associated with this feature selected from the features ensemble that describe the samples.

The critical point is the choice of these parameters. The majority of decision tree methods are based on the same strategy. At each node, an evaluation criterion is evaluated for each of the incoming features and for possible tests on these variables.

The evaluation criterion is generally based on an impurity measurement, which depends on the degree of homogeneity of the samples belonging to the node. The impurity is minimal when the samples all belong to the same class, *i.e* a pure node. On the contrary, the impurity is maximal if the samples are distributed equitably between all the classes.

Formally, the ΔI criterion seeks to maximize the impurity difference between the population P of the node and those of the populations P_l and P_r of the two child nodes.

$$\Delta I(P, P_l, P_r) = I(P) - (I(P_l) + I(P_r)) \tag{3.1}$$

where I is the impurity measurement. Only the two most commonly used impurity measurements are presented. These measurements are calculated for a population P composed of m training samples belonging to K classes. The number of samples belonging to the k-th class is noted as m_k.

The first rule is the information gain used for the induction of the decision-tree algorithm ID3 and C4.5 [48]. It consists of measuring the amount of information needed to determine the class of a sample and it is calculated based on Shannon's entropy expression.

The second measure is the *Gini* index used for induction of classification and regression trees (CART) [45]. This impurity measurement represents how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset. The *Gini* index is expressed as follows:

$$I_{Gini}(P) = 1 - \sum_{k=1}^{K} \left(\frac{m_k}{m}\right)^2 \tag{3.2}$$

Regarding the choice of a rule to decide if a node is terminal node, the simplest strategy is to decide that a node becomes a leaf if it is pure, *i.e.* all the samples contained in the node belong to the same class.

However, the construction of a tree to its maximum depth, *i.e.* without a stopping criterion, generally leads to a complex model that may perfectly adapts to the learning samples but that is unable to generalize to other data (*Overfitting*).

In order to improve the generalization ability of the constructed model and to avoid *overfitting*, it is possible to stop the construction of the tree prematurely. For instance, if:

- a maximum depth (*max_depth*) is defined by the user,
- the number of samples contained by a node is lower than a parameter (*min_samples*) defined by the user,

– the variance within the nodes does not decrease beyond a certain threshold.

Another solution is the method of *pruning*. It consists of building the tree to its maximum depth and then, to remove the non-interesting nodes.

3.3.3 From the tree to the forest

The binary decision trees described above are known to be very sensitive to *overfitting*, and to have a "weak" generalization capacity. However, their fast learning phase and the readability of their decisions make them attractive. Thus, it was proposed to build sets of binary decision trees by taking advantage of the great properties of the ensemble methods, including the improvement of the generalization capability.

In the seek for competitive performance, each decision tree should perform well (low bias, but high variance is allowed). Besides, the trees should be weakly correlated. Two trees are weakly correlated if their prediction on the same set of samples is different.

The best-known ensemble method using a set of binary decision trees is the *Random Forest - Random Input* proposed by Breiman [29], which is often called *Random Forest*. Conventionally, the label of a new observation is obtained by a majority vote on all the results of the trees built.

The RF are generally established with following features:

– to use *bootstrap* samples to construct the K decision trees that compose the final model;

– to use the principle of *random feature selection*, i.e. at each node, the partitioning criterion is evaluated only for a subset of m input features randomly drawn without replacement by means of the *Gini* criterion given by equation 3.2;

– to build the trees to their maximum depth.

The use of bootstrap samples and the principle of *random feature selection* makes it possible to diversify the trees, and therefore, to decorrelate them. Moreover, the use of a small number of features for the construction of each node makes it possible to reduce the algorithmic complexity of the RF resulting in a lower computational time.

In the used implementation (OpenCV), the tree construction is stopped prematurely if a predefined maximum depth (*max_depth*) is reached or if the number of samples within the node is lower than a parameter called *min_samples*. This variant of the initial method makes it possible: 1) to reduce the *overfitting* of the trees, and

2) to reduce the algorithmic complexity. Finally, each terminal node votes for the class present in the majority of the learning samples.

In addition, the operation of the RF allows the calculation of three particularly interesting metrics: 1) the importance of each variable, 2) the error *Out Of Bag* (OOB) , and 3) the vector of probabilities of belonging to each class for the predicted samples.

For a given tree, the learning process with *bootstrap* samples implies that some of the training samples are not used for the building of each tree. These samples not used for training the model are called OOB samples. On average, about one-third of the samples are OOB when the number of learning samples is large [37]. For each tree, the OOB samples are used to estimate the OOB error by counting the number of times than these samples are misclassified [29].

The RF is also able to give an indication of the most important features by means of the so-called *variable importance*. This information gives a knowledge about these features obtaining a better explanation of the classification result, and being able to identify those features that are superfluous and redundant. In addition, the variables of importance can be used to construct a better classifier [49].

The ensemble of trees in the RF also allows the calculation of the probabilities of belonging to the classes for a given sample. In the following section, this metric is detailed.

3.3.4 The probabilities of belonging

The RF probability vector for a sample x is denoted by $p(x)$. It is defined as the $p(x) = \{p_{c_1}(x), \cdots, p_{c_N}(x)\}$ with N the number of classes and the probability $p_{c_i}(x)$ represents the probability that sample x belongs to class C_i. The idea is to use the fact that the RF is an ensemble method. Therefore, the probability vector can be directly calculated using the set of trees in the RF [50].

As proposed by Breiman [29], the probability vector is computed by counting the number of predictions per class for all the ensemble trees. The probability $p_{c_i}(x)$ is calculated, equally on all the trees, as follows:

$$p_{c_i}(x) = \frac{1}{K} \sum_{k=1}^{K} p_{c_i}^k(x) \tag{3.3}$$

with K the number of trees in the forest, and $p_{c_i}^k(x)$ the probability that the k-th tree predicts class C_i for the sample x. The probability $p_{c_i}^k(x)$ is calculated by studying the composition of the leaf $n_k(x)$ where the sample x falls into for the k-th tree. In particular, during the building of the kth-tree, the number of samples per class that fell on the leaf $n_k(x)$ are counted in the vector $m^{n_k(x)}$. This vector is defined as: $m^{n_k(x)} = \{m_{c_1}^{n_k(x)}, \cdots, m_{c_N}^{n_k(x)}\}$. Then, the probability $p_{c_i}^k(x)$ can be

expressed as follows:

$$p_{c_i}^k(x) = \begin{cases} 1 & \text{If } argmax(m^{n_k(x)}) = C_i \\ 0 & \text{Otherwise} \end{cases} \tag{3.4}$$

For a given sample x, each tree performs its decision following Equation 3.4. This equation shows that this vote, it counts for one what means that every tree has the same vote. Then, once each tree has computed its decision, the number of predictions are counted per classes and divided by the total number of classifiers (*i.e.* K). This prediction count can be interpreted as a probability vector of belonging. Figure 3.3 presents a schematic detailing the computation of the class probability vector.

Generally, the decision of the ensemble model is given by the class that obtains more votes. This also means that the predicted label will correspond to the class with the highest probability. By means of the probability vector the user is able to obtain further information of the decision process. The interest of this work is to exploit this information to improve the classification results. Further a statistical evaluation is given for the radar and optical classification chains showing the advantages of these probabilities. Besides, part of the fusion methods proposed in this work are based on the handling of this feature.

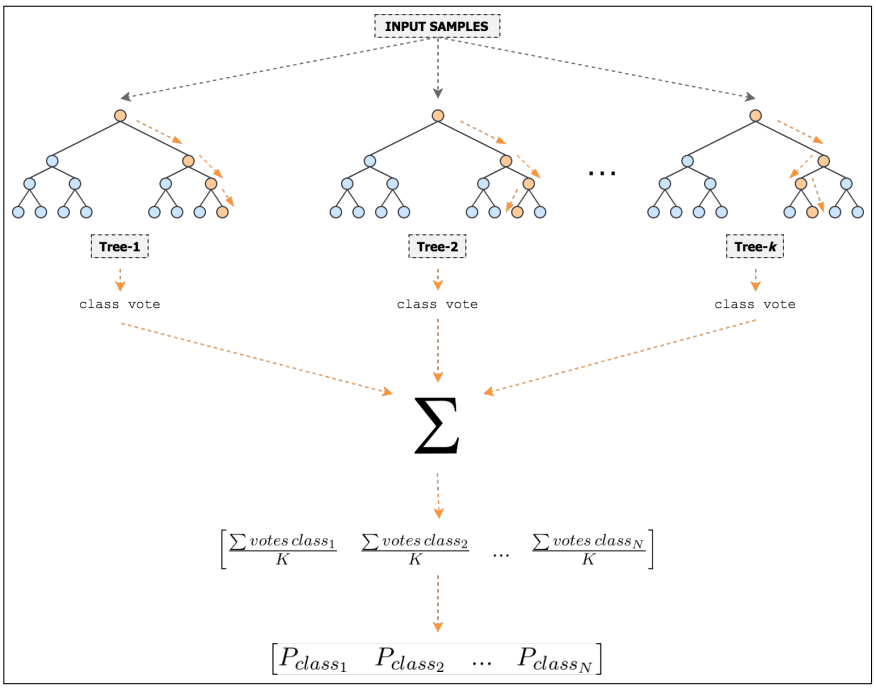


Figure 3.3: Random Forest approach to obtain the probabilities array

3.4 Classification performance evaluation

The evaluation step consists of quantifying the accuracy of the classification algorithm and, therefore, of the land cover maps produced by the classification system. This step is performed by comparing the classes of the reference data with those predicted by the classifier on the test, or validation, samples.

In order to carry out this evaluation, a double entry table called confusion matrix or contingency table is generally used. The confusion matrix is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one. Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class (or vice versa).

Consequently, the diagonal samples represent the number of test samples correctly predicted by the classification algorithm, also called true positives and true negatives samples. But not only that, this matrix reports the number of false positives and false negatives as well. This allows more detailed analysis than mere proportion of correct classifications (accuracy). Accuracy is not a reliable metric for the real performance of a classifier, because it will yield misleading results if the data set is unbalanced (that is, when the numbers of observations in different classes vary greatly).

	Predicted Class	Predicted No-Class
Actual Class	True Positive (TP)	False Positive (FP)
Actual No-Class	False Negative (FN)	True Negative (TN)

Table 3.1: Confusion matrix for a binary classification problem.

From this confusion matrix, it is possible to calculate a set of metrics characterizing the performance of the classification algorithm used. The test samples used must reference the class in the field so as not to bias the results. In practice, the test samples are considered perfect (gold standard).

In a first part, the evaluation metrics in the case of a multi-class classification are described. A second part presents a way of estimating the real error committed by the classification system by relying on the confidence interval.

3.4.1 Multi-class classification evaluation

In the context of a K-class classification problem, the confusion matrix C is defined as follows:

$$C = \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1j} & \dots & c_{1K} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ c_{i1} & c_{i2} & \dots & c_{ij} & \dots & c_{iK} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ c_{K1} & c_{K2} & \dots & c_{Kj} & \dots & c_{KK} \end{pmatrix}$$

For a classification that studies K classes, the confusion matrix C will be an $K \times K$ matrix with each element c_{ij} being the number of pixels predicted to belong to the class i when the actual class is of type j. The test samples correctly predicted by the classification algorithm are on the diagonal of the confusion matrix (coefficients c_{ij}) as mentioned before.

The simplest metric derived from the confusion matrix is the rate of good classifications, the so-called Overall Accuracy (OA) which is calculated as the number of correctly predicted test samples (trace of the confusion matrix) divided by the total number of test samples:

$$\text{OA} = \frac{1}{N} \sum_{i=1}^{K} c_{ii} \tag{3.5}$$

where $N = \sum_{i=1}^{K} \sum_{j=1}^{K} (c_{ij})$ is the total number of test samples.

In addition to the OA, it is common to calculate the Kappa coefficient which is a more robust measure than simple percent agreement calculation, as Kappa index takes into account the possibility of the agreement occurring by chance:

$$\text{Kappa} = \frac{\text{OA} - p_e}{1 - p_e} \tag{3.6}$$

where p_e is the hypothetical probability of chance agreement. Hence, p_e is determined using the observed data to calculate the probabilities of each observer randomly seeing each category and is defined as $p_e = \frac{1}{N^2} \sum_{i=1}^{K} (\sum_{j=1}^{K} c_{ij})(\sum_{j=1}^{K} c_{ji})$.

If the raters are in complete agreement then $K = 1$. If there is no agreement among the raters other than what would be expected by chance (as given by p_e), $K \approx 0$.

Global metrics such as OA and Kappa are often insufficient to measure the quality of the classification, especially in the case where the number of test samples per class is very unbalanced. Indeed, these measures do not take into account the distribution of classes. Consider the real case of mapping *pines versus oak trees*, where pines are much more present than oaks. For example, 95 trees are pines and 5 trees are oaks. If the classification algorithm decides that all trees are pines, then the OA will be 95%. In this case, the value of OA does not make it possible to highlight that no oak is detected by the classification algorithm used.

In order to take into account differences in performance between classes, metrics by class are used. The *precision* and the *recall* defined for the i-th class are commonly used:

$$precision_i = \frac{c_{ii}}{\sum_{j=1}^{K} c_{ji}} \tag{3.7}$$

$$recall_i = \frac{c_{ii}}{\sum_{j=1}^{K} c_{ij}} \tag{3.8}$$

In other words, the precision of a class corresponds to the percentage of correctly predicted samples in this class compared to all the predictions made for this class, while the recall represents the percentage of correctly retrieved samples over the total amount of reference data for that class. Depending on the application, only one of the two measures may be of interest.

In the previous example, if it is important to detect all the oaks, a strong *recall* value will be preferred, even if you detect too many pine trees as oaks (low *precision*). In contrast, if the objective is to be certain that the trees detected as oaks are oaks, the *precision* will have to be maximized, even if some are missing (weak *recall*).

For most applications, a compromise between accuracy and recall is usually desired. Then, it is possible to combine both measurements into a single one named the F-Score (or F-1) defined as their harmonic mean:

$$F\text{-}Score_i = 2\frac{recall_i \times precision_i}{recall_i + precision_i} \tag{3.9}$$

The higher the F-score (that takes value between 0 and 1), the better is the performance of the classifier in detecting a given class.

3.4.2 Statistic evaluation

All these measures are obtained from classifications that are performed on small samples of all the available reference data. Because of this and also because of the intrinsic randomness of the classification algorithms, those measures are subject to uncertainties.

These uncertainties can be evaluated thanks to the calculation of the *95% Confidence Interval* . It simply gives the interval where the real value of the measure has a 95% chance to be found. The confidence interval is measured for n different random draws for which the classification outcome gives an estimation X of a given measure. Therefore, its empirical mean $\bar{X}$ and standard deviation σ_X can be calculated for those n random draws. They satisfy:

$$X = \bar{X} \pm t_{95}(n)\frac{\sigma_X}{\sqrt{n}}i, \tag{3.10}$$

where $t_{95}(n)$ is the 95% quantile associated to the Student t-distribution commonly used when the number of draws is smaller than 30. For this work, random draws are used for three-tile zone datasets. The corresponding quantiles are well approximated by $t_{95}(10) \simeq 1.833$.

Chapter 4

Description of the classification chains

This project has been carried out as part of the SENSAGRI [1] project. The interest of this project is to exploit the synergies between radar and optical data provided by the Sentinels missions for agricultural mapping purposes. The final product should be able to obtain land cover maps at dealing with huge amount of data, different kinds of agricultural systems and different eco-climatic areas. Hence, the SENSAGRI project implies different constraints and requirements. The classification system will operate with the minimum human intervention but with enough flexibility in order to deal with the different operational conditions. In order to carried that out, four approaches have been taken into account. On the one hand, two single classification chains have been performed which it means that radar and optical information is separately classified. On the other hand, all the available remote sensing data is used as input data for the classifier model. In other words, information provided by Sentinel-1 and -2 is integrated at pixel level and classified in one single processing chain. The interest of this work is to present a fusion strategy, at decision-level, able to outperform the former approaches.

In this chapter the classification chains for the different classification strategies are presented. The different blocks that composed a classification system are also briefly described. Then, this chapter presents the following structure. Firstly, the channels extraction stage is commented. Secondly, a description of the proceedings for the tuning of the supervised model is given. Finally, the different classification system are presented.

4.1 Channels Extraction

As explained in Section 2.3, satellite images are composed by different information. For instance, different spectral bands per acquisition are contained in optical satellite

[1] http://www.sensagri.eu/

images. Besides, pre-processing tasks have been carried out in these images. Hence, once the satellite images are processed, the channels extraction step is performed. In this block, the bands composing each image are extracted and given to the classifier model.

The main optical channels per acquisition correspond to the ten spectral bands detailed at Table 2.2. Besides, the addition of some spectral indices has been proposed to add to the input data set but the use of this configuration is out-of-scope of this project. In contract, the main radar features correspond to the two polarizations (VH and VV). Similarly, additional features has been proposed as the ratio of polarization $\frac{VH}{VV}$ to minimize the effect of soil moisture and roughness.

As commented in Section 2.1, the interest of this project is to exploit the Sentinels time series and, therefore, to perform multi-temporal classification. For this reason, the number of extracted bands will increase as the end of the season comes. Thus, 330 optical bands are extracted and classified at the end of the agricultural season. Accordingly, 168 bands are extracted for the radar case.

4.2 Learning the RF classification model

The goal of the learning step is to construct an accurate classification model which is used to predict unlabeled samples. For this case, the classification model is applied on the complete image times series to obtain the final land cover maps. The learning strategy presented here can be divided in three steps:

1. The construction of two independent learning and testing data sets from the reference data.
2. The definition of a random sampling strategy.
3. The choice of the Random Forest parameters.

4.2.1 Sampling the reference data

Concerning the reference data, it corresponds to a vector file composed of reference polygons. More information about this data set can be found in Section 2.2. The purpose of these reference polygons is twofold: (1) to validate and (2) to learn the supervised classifier. In order to meet the requirements, different strategies have been defined for the reference data.

Splitting polygons for training and testing

The reference data is composed of crop and "non-cropland" polygons. The goal is to split these data into two disjoint subsets: the training set and the validation

set. These sets are composed of polygons, not individual pixels. The strategy to split the data set is carried out by a random splitting algorithm. This algorithm involves three vectors of polygons: 1) The *reference polygons* vector, 2) the *training polygons* vector and the 3) *validation polygons* vector. Besides, a ratio, called *sample ratio*, containing the ratio between the number of training and validation polygons per class is recursively computed. By means of these parameters, the algorithm performs a random sampling without replacement of the polygons of each class with probability p for the training set and $1-p$ for the validation set. In algorithm 1 it can be seen the flow of this strategy.

Algorithm 1: Splitting reference data

Data: $reference\_polygons$, $sample\_ratio$, $list\_of\_class\_label$
Result: $training\_polygons$, $validation\_polygons$

```
begin
    training_polygons ← 0;
    validation_polygons ← 0;
    for cl ∈ list_of_class_label do
        for poly ∈ reference_polygons do
            if class_of(poly) = cl then
                p ← random(0, 1);
                if p ≤ sample_ratio then
                    add poly to training_polygons
                else
                    add poly to validation_polygons
                end
                remove poly from reference_polygons
            else
            end
        end
    end
end
```

The resulted training and validation polygons data sets are composed by four different fields containing important information for the classification chain:

(i) Crop (with value 1) or no-crop (with value 0)

(ii) Crop type ID

(iii) Class name

(iv) Origin of the reference data

This splitting procedure is repeated 10 times ($nbruns$) with different random draws from training and validation samples in order to statistically evaluate the results by computing confidence intervals allowing to reduce variability of the data random selection.

Random sampling strategy on learning data set

In order to guarantee that samples used to learn the classification algorithm fulfill the required conditions, a sampling strategy was defined. It consists in randomly selecting samples from the *training polygons* vector file. This step is required since *training polygons* can contain: (1) too many reference polygons covering a very large geographical area (having an important landscape diversity) or (2) unbalanced class distributions (minority classes). Therefore, this sampling strategy allows to obtain a balanced training set that meets with a minimum and maximum training sample size per class.

4.2.2 Training strategy

As explained in Section 3.3, Random Forest classifier is a well-known ensemble learning method that grows *nbtrees* classification trees. To classify a new sample, each tree gives a classification and we say the tree "votes" for that class. Finally, the forest chooses the classification having the most votes (over all the trees in the forest).

The RF trees are built by randomly selecting at each node a subset of input variables (denoted by m). Several works have proved that the optimal value of m corresponds to the square root of the number of input variables [37]. The construction of a tree is recursively done by splitting the RF node (using the random m variables) into more homogeneous nodes.

Ideally, this random selection is repeated recursively on each derived sub-set until the node contains very similar samples, or when the splitting no longer adds value to the predictions. For implementation purposes, the tree building can be stopped when a maximum depth (*max_depth*) is reached, or when the number of samples at the node is below a *min_samples* threshold.

Therefore, the use of Random Forest needs the tune of the four parameters described at Table 4.1. These parameters has been tuned by following the recent work [20].

Notation	**Definition**
nbtrees	This is the number of trees you want to build to predict (100)
m	These are the maximum number of features allowed to try in individual tree ($\sqrt{p}$)
max_depth	The maximum depth of the tree (25)
min_samples	The minimum sample leaf size (25)

Table 4.1: Random Forest parameters (p = the number of input variables)

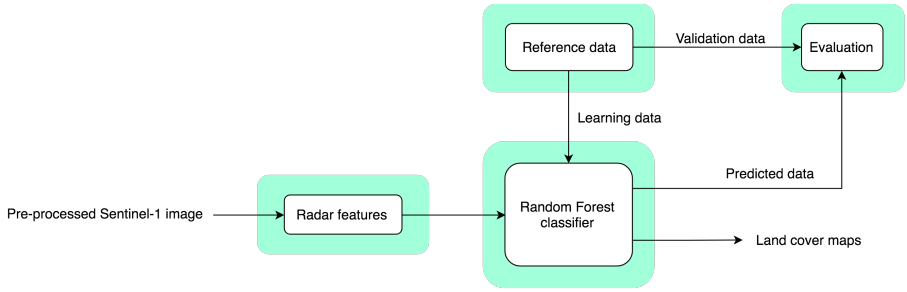
4.3 Supervised classification systems

In this section the classification strategies used for the purposes of this work are presented. On one hand, the input data needed for this project were introduced in Chapter 2, there, the satellite time series used were described as well as the pre-processing tasks carried it out in order to obtain a suitable data to classify. On the other hand, the supervised classification process has been detailed in Chapter 3, where the Random Forest algorithm were presented and explained. Besides, the features extraction, the processing of the reference data and the tuning of the RF model have been given in this chapter. Besides, the mapping chains should fulfill the following requirements: 1) the satellite data set is composed of all the possible optical and (or) radar images, 2) the input data shall be pre-processed by the steps described at Section 2.3 and 3) the legend describing land cover classes is "fixed" along the agricultural season.

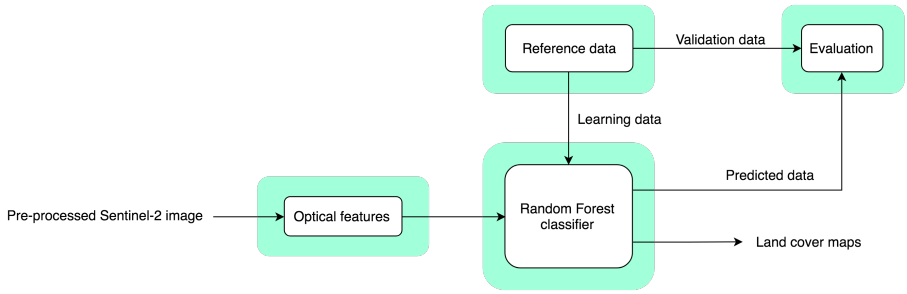
This last section aims to present different schemes where put into operation all these blocks and requirements. For these reason four different approaches are given. The first three strategies are former developments to this work. Therefore, the purpose of the last strategy is to outperform them.

Single classification system

The first presented classification system is illustrated in Figure 4.1, which must be interpreted from left to right. This figure can be easily divided in three main tasks (yellow rectangles) : the optical (or radar) feature extraction, the classification and the evaluation step. The satellite input data correspond to the pre-processed optical Sentinel times series. Concerning the reference data, it corresponds to a vector file composed of reference polygons. Also, the figure shows the two uses of the reference data: (1) to validate and (2) to learn the supervised classifier. Finally, regarding the supervised classification step, as explained before, corresponds to the RF classification algorithm.



(a) Radar supervised classification system



(b) Optical supervised classification system

Figure 4.1: Single classification strategy

As it can be seen, this strategy only exploits the RS data from one source. In Chapter 6, the results of these classifiers are given and analyzed.

Radar and optical integrated data classification system

In Figure 4.2 the second classification system is presented. This approach is based on the use of all the available radar and optical data as input data. From the fusion viewpoint, it can be seen as fusion at pixel level.

This strategy implies a high dimensionality since it gathers all dates and bands from radar and optical satellites acquisitions. As mentioned before, Sentinel-1 and -2 have different temporal resolutions. In Figure 2.1 it was shown this assessment. Therefore, an algorithm to perform the temporal integration was implemented. This algorithm considers all the available dates from optical and radar satellites acquisitions. Also, they do not need to be synchronized. In Table 4.2 an example of this temporal integration is given for a better understanding.

Dates	RS data	Bands per date	Number of total bands
1	radar	2	2
2	radar	2	2
3	optical	10	4
4	radar	2	14
5	optical	10	26
6	optical	10	36

Table 4.2: Example of the temporal integration.

As it can be seen in this table, the fourth column of the table shows the total number of bands sent to the classifier at a given date. Then, the algorithm takes all the available dates and concatenates the available bands.

Figure 4.2 shows a schematic of this classification strategy. In this flowchart, it can be seen how the ensemble of features is given to the RF model for the learning and classification steps.

In Chapter 7 the performance of this approach is shown in order to compare it with the other classification strategies.

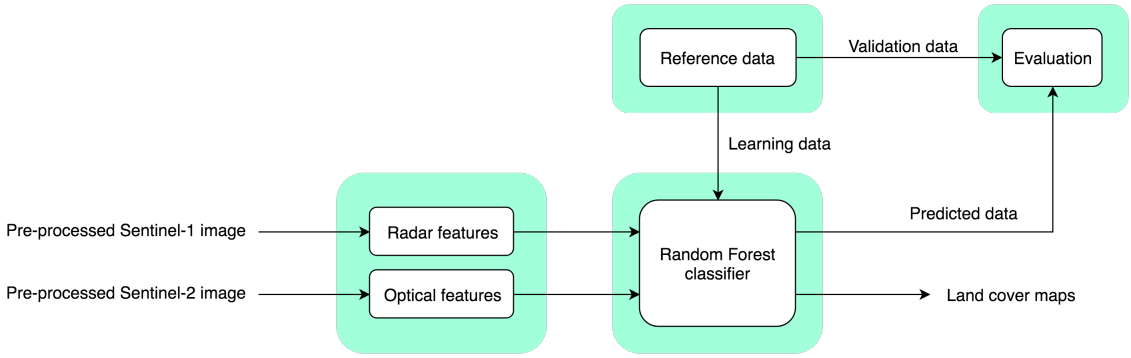


Figure 4.2: Classification system that involves the integration of all the available radar and optical input data.

Fusion classification system

The fusion at decision-level allows the classification system to parallelized and reduce the computational burden that implies the classification of high-dimensional data. Figure 4.3 presents this approach. As it can be seen, the radar and optical classification are performed separately. The same learning data and training parameters are used for both models. The main idea of this strategy is not to rely in only one classifier in the same way as ensemble methods do. Next chapter presents a detailed description of the proposed fusion methods for this approach.

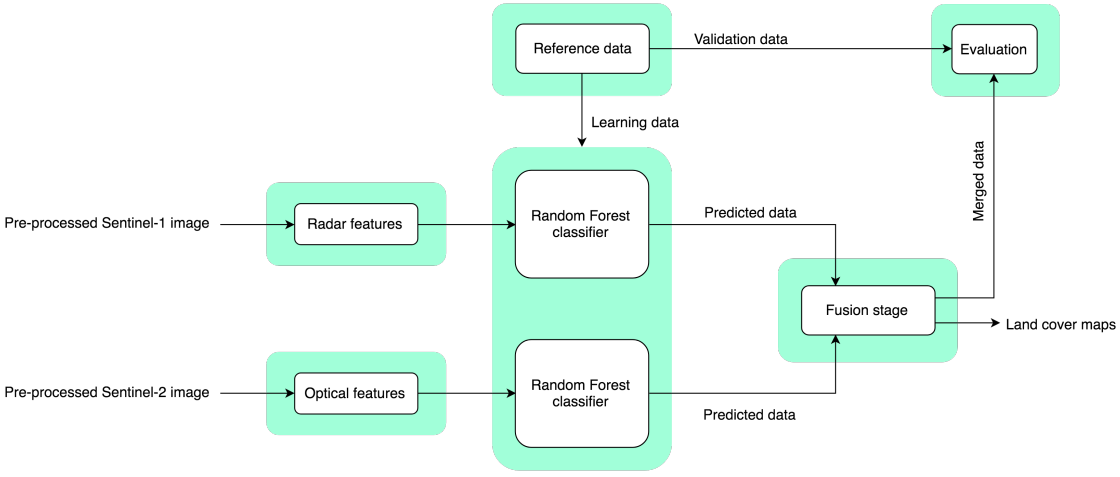


Figure 4.3: Fusion at decision-level classification strategy.

Chapter 5

Fusion of classifications

Data fusion is a powerful tool to combine data derived from disparate sources such that, the resulting information has less uncertainty than would be possible when these sources were used individually. In our case, this works proposes different strategies to exploit optical and radar satellites data to improve land cover mapping services. The probabilistic outputs of RF model allows us to propose different fusion strategies. Those fusion methods are based on the idea that different classifiers typically express their opinions in different ways. Besides, many researchers have worked on combining the radar and optical data, which has lead to improve their overall classification accuracy results [51] [52].

The objective of this chapter is to describe a set of fusion classifiers techniques in order to exploit the results of the Sentinel-1 and Sentinel-2 satellite images classifications.

5.1 Fusion data approaches

Methods of data fusion can be grouped into three categories depending on the level at which the integration is performed: 1) pixel-level fusion (or data fusion), 2) feature-level fusion and 3) decision-level fusion [21]. The first category refers to the combination of the original image pixels and the simplest approach is to concatenate the data from the different sensors as if they were measurements from one single sensor [53]. The second category is based on combining features extracted from the individual datasets [54]. In contrast, decision fusion requires preliminary analysis of the different datasets, (*e.g.*, the separate classifications of optical and SAR data), after which outputs are combined to generate a final result.

Different combination strategies implying data fusion at decision level have been proposed for this research in order to obtain an integrated classifier. The idea is not to rely on a single decision making scheme. Then, all the designs are used for decision making by combining their individual opinions to derive a consensus decision. Therefore, the results of the single classifiers were then combined into an

ensemble using five methods: a maximum confidence rule, a median rule, a method based on Bayes' rule and two methods based on Dempster-Shafer theory [55].

Moreover, this works aims to exploit the information provided by the RF probabilities in order to improve the fusion at decision level. As mentioned before in Section 3.3, the class probability vector allows us to obtain a better understanding of the decision process. Therefore, the interest of this work is to combine the advantages of these probabilities and the complementary information from several classifiers.

5.2 The Dempster-Shafer fusion

Mathematical theory of evidence was first introduced by Dempster in the 1960's, and later extended by Shafer [55]. This theory, which allows to represent both imprecision and uncertainty, appears as a more flexible and general approach than the Bayesian one. Another of its advantages is its ability to consider not only single (or individual) classes, but also unions of classes. Applications were developed in medical imaging, object detection, and remote sensing classification [56].

The Dempster-Shafer (DS) theory of evidence, also known as the theory of belief functions, is a tool for representing and combining evidence. Being a generalization of Bayesian reasoning, it does not require probabilities for each question of interest, but the belief in a hypothesis can be based on the probabilities of related questions. Contributing to its success is the fact that the belief and the ignorance or uncertainty concerning a question can be modelled independently.

The Dempster-Shafer Theory

The Dempster-Shafer theory [56] starts by assuming a universe of discourse, or frame of discernment, consisting of a finite set of mutually exclusive atomic hypotheses $\Theta = \{\theta_1, ..., \theta_q\}$. In image classification applications, Θ is the set of hypotheses about pixel class. Dempster–Shafer theory allows to consider any subset of Θ. In the following, Let denote 2^Θ the set of the subsets of Θ. Applied to classification problems, it means that not only single classes (also called singletons) but also any union of classes can be represented. In the following, hypotheses about singletons and hypotheses about unions of classes are respectively called simple hypotheses and compound hypotheses.

By extension of the notations of the set theory, inclusion, intersection, and union between two hypotheses A and B are defined and denoted as follows:

$$\text{for a given event } x: \quad \forall A \in 2^{\Theta},\ \forall B \in 2^{\Theta}$$

$$\begin{cases} A \subseteq B \Leftrightarrow \text{if } A \text{ is true, then } B \text{ is true} \\ (A \cap B) \text{ is true} \Leftrightarrow \text{if } A \text{ is true and } B \text{ is true} \\ (A \cup B) \text{ is true} \Leftrightarrow \text{if } A \text{ is true or } B \text{ is true} \end{cases}$$

Representation of Evidence

The Dempster–Shafer evidence theory provides a representation of both imprecision and uncertainty through the definition of two functions: plausibility (Pls) and belief (Bel), which are both derived from a mass function (m). Then a function m is defined for each element A of 2^{Θ}, such that the mass value $m(A)$ belongs to the [0, 1] interval and is called a basic probability assigment (bpa) if it satisfies:

$$m : \begin{cases} m(\emptyset) = 0 \\ \sum_{A \in 2^{\Theta}} m(A) = 1 \end{cases} \tag{5.1}$$

where $\emptyset$ is the empty set.

The belief in $m(A)$ represents the ignorance, which can not be subdivided among the subsets of A. Then, when the mass affected to a compound hypothesis $A \cup B$ is nonzero, it means that there is an option not to make the decision between A or B but rather leave the pixel in the $A \cup B$ class. In particular, assigning a non null mass to Θ allows to not classify some pixels, for which there is a global ignorance.

The belief and plausibility functions, derived from m, are respectively defined from 2^{Θ} to [0, 1]:

$$Bel(A) = \sum_{B \subseteq A} m(B) \tag{5.2}$$

$$Pls(A) = \sum_{B \cap A \neq \emptyset} m(B) \tag{5.3}$$

These two functions, which have been sometimes referred to as lower and upper probability functions, have the following properties:

$$\begin{cases} Bel(A) \leq Pls(A) & (5.4) \\ Pls(A) = 1 - Bel(\overline{A}) & (5.5) \end{cases}$$

where $\overline{A}$ is the complementary hypothesis of A: $A \cup \overline{A} = \Theta$ and $A \cap \overline{A} = \emptyset$.

In the case of Bayes theory, uncertainty about an event is measured by a single value (probability) and imprecision about uncertainty measurement is assumed to be null. In the case of Dempster–Shafer theory, the belief value of hypothesis may be interpreted as the minimum uncertainty value about, and its plausibility value, which is also the "unbelief" value of the complementary hypothesis $\overline{A}$ [see (5.5)], may be interpreted as the maximum uncertainty value of A. Thus, uncertainty about A is represented by the values of the interval [Bel(A), Pls(A)], which is called the "belief interval" and the length of this belief interval gives a measurement of the imprecision about the uncertainty value.

In simple terms, the $Bel(A)$ represents the minimum trust in A because of the supporting subsets B. Looking at the definition, it can be noticed that there is a one-to-one correspondence between the belief function and the basic probability assigments. If A is an atomic hypothesis, $Bel(A) = m(A)$. To get an intuitive understanding, one can consider a basic probability assignment a generalization of a probability density function and a belief function a generalization of a probability function.

The mass of Evidence

The combination of classifiers is based primarily on the consideration of the errors of individual classifiers. The errors of each classifier are usually recorded in the confusion matrix:

$$M^j = \begin{pmatrix} n_{11} & n_{12} & \dots & n_{1j} & \dots & n_{1N} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ n_{i1} & n_{i2} & \dots & n_{ij} & \dots & n_{iN} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ n_{N1} & n_{N2} & \dots & n_{Nj} & \dots & n_{NN} \end{pmatrix}$$

where N is the number of classes. The row i corresponds to class C_i and column j corresponds to the class determined by classifier j.

As explained, this matrix can be considered as a *priori* knowledge on the performance of the classifier. The diagonal elements are the percentages of matches between classes reconstructed by the classifier and reference classes. There is a confusion matrix for each classifier. The masses of evidence will be determined by:

$$m_j(\{C_k\}) = \frac{n_{kk}}{\sum_{i=1}^{N} n_{ki}} \tag{5.6}$$

Other methods were studied but the choice of this method was motivated by the fact that it takes into account the recognition rate per class.

Evidence Combination

Dempster–Shafer theory provides a method to combine the previous measures of evidence of different sources. If m_i is the basic probability assignment provided by source $\sum^i (1 \leq i \leq p)$, the combination: $m = m_1 \oplus \cdots \oplus m_p$, also called orthogonal sum, is defined, according to the Dempster's combination rule [55], by

$$\begin{cases} m(\emptyset) = 0 & (5.7) \\ \text{if } K \neq 1, m(A) = \dfrac{\sum_{B_1 \cap \cdots \cap B_p = A} \prod_{1 \leq i \leq p} m_i(B_i)}{1 - K} & (5.8) \\ \text{where } K = \displaystyle\sum_{B_1 \cap \cdots \cap B_p = \emptyset} \prod_{1 \leq i \leq p} m_i(B_i) & (5.9) \end{cases}$$

From (5.7), it might be seen that $K (K \in [0, 1])$ represents the mass which would be assigned to the empty set, after combination, in the absence of normalization [division by $(1 - K)$ in (5.8)]. Thus, K is often interpreted as a measure of conflict between the different sources and it is introduced in (5.8) as a normalization factor. The larger K is, the more the sources are conflicting and the less sense has their combination. Finally, the orthogonal sum does not exist when K is equal to 1. In this case, the sources are said to be totally or flatly contradictory, and it is no longer possible to combine them.

Decision Rule

Having computed the mass, plausibility and belief values for each simple and compound hypothesis of the multi-source model, we need a criterion, which is called "decision rule", to decide which hypothesis is the more "realistic". Nowadays, the choice of this criterion remains application dependent. The three most popular decision rules are ([57], [56]): 1) the maximum of plausibility, 2) the maximum of belief, and 3) the maximum of belief without overlapping of belief intervals. The maximum of plausibility has been used by some authors [58]. The maximum belief over the simple hypotheses is the simplest and more used. Rule 3), also called absolute decision rule, is very strict. Other rules such as $max\{Bel(A) + Pls(A)\}$ [which may also be written $max\{Bel(A) - Bel(\overline{A})\}$] are compromises.

Among the existing rules of decision, for this study the decision of the maximum belief has been considered. Hence, for each possible assigned class C_i (with $i \leq N$), the belief in that class $Bel(C_i)$ is computed as detailed in Equations 5.3 and 5.9.

Therefore, the assigned class to a given pixel x after the fusion step is computed as follows:

$$C_{fusion}(x) = \arg\max_{C_i} \{Bel(C_i)\} \quad i = 1, \cdots, N \tag{5.10}$$

The Dempster-Shafer fusion has presented the merge of the decisions from several classifiers. Therefore, although the accuracy of each classifier is considered, the merge is based solely on the output label. The following sections aim to present a different approach where the fusion involves also the probabilities of belonging obtained from the RF algorithm for a given classifier.

5.3 Bayesian Belief integration

The following strategy aims to exploit the class probabilities obtained from the RF model. The interest of this method is show how the class probabilities are able to provide a better metric of the classifier. Several methods of merging consider the accuracies of each classifier based on metrics computed from the confusion matrix. Those *a posteriori* approaches give only a general understanding of how the classifier is performing. For these reasons, the purpose of the probabilities management is to improve the fusion of decisions. Besides, this enables the fusion methods a better knowledge since the probabilities are obtained for each pixel and RF model.

This fusion uses Bayesian methodology to provide a belief measure associated with each classifier output and eventually integrates all single beliefs resulting in a combined final belief. The quality of this fusion depends on how the posterior probabilities are estimated and the diversity of used classifiers.

This method is defined as follows. Consider a sample being classified as C_j (with $1 \leq j \leq N$) given the observation X_{obs}, where X_{obs} corresponds to the event that the classifier e_k has assigned the class C_j to sample x. Therefore, the posterior probabilities for the event X_{obs} can be denoted as:

$$P(x \in C_i \mid e_k(x) = C_j) \qquad i = 1, \ldots, N \tag{5.11}$$

where i denotes all the possible classes of belonging and $e_k(x) = C_j$ corresponds to the event X_{obs}. As explained before, this work aims to exploit the probabilities obtained from the RF model. For this reason, these probabilities (see Equation 3.3), can be regard as the knowledge of expert e_k, being $P(x \in C_i \mid e_k(x) = C_j)$ the probabilities estimated for the RF model.

Based on the occurrence of the event $e_k(x) = C_j$, the expert expresses its beliefs with uncertainty on each of the N mutually exclusive propositions $x \in C_i$. Then, this uncertainty is defined by a real value called *belief*. The belief value, given the decision of a classifier $e_k(x) = C_j$ with respect to an example x, can be defined as:

$$bel(x \in C_i \mid e_k(x)) = P(x \in C_i \mid e_k(x) = C_j) \qquad i = 1, \ldots, N \tag{5.12}$$

Consider D different experts $e_1, \cdots, e_D$, with $(D \times N)$ probabilities of belonging $P(C_1 \mid e_k), \cdots, P(C_N \mid e_k)$, used over the same observation x. Then, D events $e_k(x) = C_{j_k}$, with $k = 1, \cdots, D$, will occur. Each event will supply its own set of

$bel(x \in C_i \mid e_k(x))$, $i = 1, \cdots, N$. Now we need to find a way to integrate these beliefs to give a combined value, given by:

$$\begin{aligned} bel(C_i) &= bel(x \in C_i) \mid e_1(x), \ldots, e_D(x)) \\ &= P(x \in C_i) \mid e_1(x) = C_{j_1}, \ldots, e_D(x) = C_{j_D}) \\ &= \frac{P(e_1(x) = C_{j_1}, \ldots, e_D(x) = C_{j_D} \mid x \in C_i) P(x \in C_i)}{P(e_1(x) = C_{j_1}, \ldots, e_D(x) = C_{j_D})} \end{aligned} \tag{5.13}$$

where $i = 1, \cdots, N$. If the classifiers perform independently of each other, then the events $e_k(x) = C_{j_k}$, with $k = 1, \cdots, D$, will be independent of each other and:

$$\begin{aligned} bel(C_i) &= \frac{\prod_{k=1}^{D} P(e_k(x) = C_{j_k} \mid x \in C_i)}{\prod_{k=1}^{D} P(e_k(x) = C_{j_k})} P(x \in C_i) \\ &= \frac{\prod_{k=1}^{D} P(e_k(x) = C_{j_k}, x \in C_i)}{\prod_{k=1}^{D} P(e_k(x) = C_{j_k}) \prod_{k=1}^{D} P(x \in C_i)} P(x \in C_i) \\ &= \frac{\prod_{k=1}^{D} P(x \in C_i \mid e_k(x) = C_{j_k})}{P(x \in C_i)^D} P(x \in C_i) \end{aligned} \tag{5.14}$$

where $P(x \in C_i \mid e_k(x) = C_{j_k})$ may be estimated by Eq. 5.11. Finally, an estimation of $P(x \in C_i)$ is established in order to reduce the computational burden. The value of the $bel(C_i)$ is estimated by:

$$bel(C_i) \approx \frac{\prod_{k=1}^{D} P(x \in C_i \mid e_k(x) = C_{j_k})}{\sum_{l=1}^{N} \prod_{k=1}^{D} P(x \in C_l \mid e_k(x) = C_{j_k})} \tag{5.15}$$

This substitution will ensure that $\sum_{i=1}^{N} bel(C_i) = 1$, condition required since the events $x \in C_i$, $i = 1..N$ are mutually exclusive and exhaustive.

As a result of this ensemble method, from the outputs of the classifiers, we obtain a *belief* about the event x depending on the combination of decisions of the two classifiers.

More precisely, and accordingly to the proposed fusion scheme in Section 4.3, *belief* expression will be estimate as follows:

$$bel(x \in C_i) \approx \frac{P(x \in C_i \mid S_1(x) = C_{j_{S_1}}) P(x \in C_i \mid S_2(x) = C_{j_{S_2}})}{\sum_{l=1}^{N} P(x \in C_l \mid S_1(x) = C_{j_{S_1}}) P(x \in C_l \mid S_2(x) = C_{j_{S_2}})} \tag{5.16}$$

where $N = 16$ classes, $C_{j_{S_1}}$ and $C_{j_{S_2}}$ the output labels predicted by the radar and optical classifiers, respectively.

Finally, the merged output label will be obtained from the maximum *belief*:

$$C_{fusion}(x) = \arg\max_{C_i}\{bel(x \in C_i)\} \quad i = 1, \cdots, N \tag{5.17}$$

As it may be seen, this fusion approach performs a product operation between the probabilities a *posteriori* from the different classifiers. The use of the RF probabilities may imply important improvements on the classification performance. However, this approach considers all the classifiers as equals giving them the same weight.

5.4 Maximum Confidence fusion

The Maximum Confidence method is the most straightforward technique and it was developed as a proof-of-concept for the fusions based on the RF probabilities.

Following the previous section, the event of a sample x being classified as class C_j by the classifier e_k can be denoted as $e_k(x) = C_j$. Also, the probability that x belongs to the class C_i predicted by e_k is defined as:

$$P(x \in C_i \mid e_k(x) = C_j) \qquad i = 1, \ldots, N \tag{5.18}$$

where i denotes all the possible classes. This strategy presents a fusion where the label predicted with higher "confidence" is assigned as the final output label. For this case, the "confidence" of a classifier will be denoted as the highest probability value obtained. Then, the fused label will be determined by the next statement:

$$C_{fusion}(x) = \arg\max_{C_i}\{P(x \in C_i \mid e_k(x) = C_{j_k}, \quad 1 \leq k \leq D)\} \tag{5.19}$$

where k denotes all the possible classifiers. But more precisely, and accordingly to the proposed fusion scheme, the output label can be estimated by:

$$C_{fusion}(x) = \arg\max_{C_i}\{P(x \in C_i \mid S_1(x) = C_{j_{S_1}}), P(x \in C_i \mid S_2(x) = C_{j_{S_2}})\} \tag{5.20}$$

where $C_{j_{S_1}}$ and $C_{j_{S_2}}$ are the output labels predicted by the radar and optical classifiers, respectively.

This strategy aims at exploiting the class probability vector. Therefore, it can be defined as a maximum operator between the highest probability obtained for each classifier. In contrast to Dempster-Shafer method, this approach, and the rest of probabilistic approaches, is not considering the classifiers accuracy.

5.5 Modified Dempster-Shafer fusion

The modified Dempster-Shafer method is proposed in this work. This approach aims to merge the concepts detailed in Section 5.4 and the advantages of the Dempster-Shafer Theory. For this reason, a modification of the Theory of Evidence is presented here in order to exploit the advantages of the probabilities obtained from the RF models.

The Dempster-Shafer presented in Section 5.2 is a widely used decision method but it has a heavy dependency on the output labels of each classifier. It means that the fusion step only can assign a class that it has been previously assigned by optical or radar single classifier. In fact, the fusion criterion of the Dempster-Shafer theory only takes into account the initial label and the accuracy of this label. Therefore, given an input class, the same fusion criterion is applied for all the predicted samples. This is not good because since the uncertainty for each pixel is different.

The presented fusion method aims to fusion the different decisions by taking into account the uncertainty that involves each sample.

Consider $P(x \in C_i) \mid e_k(x) = C_j)$ the probability of the sample x of belonging to the class C_i for a given classifier e_k and C_j the predicted label for this classifier. Also, consider the belief computed by the Dempster-Shafer fusion $bel_{ds}(C_i)$ (see Section 5.2). Hence, a redefinition of the belief is detailed. The belief of the presented fusion can be denoted as:

$$bel_{mds}(x \in C_i \mid e_k(x) = C_j) = bel_{ds}(C_i)P(x \in C_i \mid e_k(x) = C_j) \tag{5.21}$$

where $i \in \{1 \cdots N\}$.

In contrast to the Dempster-Shafer approach, for this case the *belief* will be calculated for each classifier. Also, the fusion step shall be computed for each pixel increasing the computational burden. Hence, the output label will correspond to the maximum *beliefs* computed for each classifier for a given pixel x. The final merged class can be defined for our case as:

$$C_{fusion}(x) = \arg\max_{C_i}\{bel_{mds}(x \in C_i \mid S_1(x) = C_{jS_1}), bel_{mds}(x \in C_i \mid S_2(x) = C_{jS_2})\} \tag{5.22}$$

Therefore, once the belief bel_{ds} is computed by means of the Dempster-Shafer theory, those values are weighted with the highest probabilities corresponding to the predicted classes. In other words, for a given sample, the *belief* bel_{mds} in the predicted classes is recomputed. Hence, the purpose of this method is to take advantage of the RF probabilities and the classifiers weights computed by the Dempster-Shafer development.

5.6 Median fusion rule

The Median Rule accomplish one of the most straightforward fusion methods but it does not imply a worse performance. As in the case of the last methods, this fusion strategy is based on the class probabilities presented in Section 3.3. The goal of this approach is to compute an averaged probability vector. Hence, given the set of class probabilities for each different classifier, the approach consists into compute the mean between all these probabilities. Once the mean is calculated, it is obtained a new probability vector for all the possible classes. These new probabilities are called *beliefs* since they show the certainty for a given class of all the classifiers.

Consider $P(x \in C_i) \mid e_k(x) = C_{j_k})$ the probability of the sample x of belonging to the class C_i for a given classifier e_k. Then, the merge rule may be defined as follow:

$$bel(x \in C_i) = \frac{1}{D} \sum_{k=1}^{D} P(x \in C_i \mid e_k(x) = C_{j_k}) \quad i = 1, \cdots, N \tag{5.23}$$

In our case, considering $e_1 = S_1$ as the radar classifier and $e_2 = S_2$ as the optical classifier, the *belief* could be defined as:

$$bel(x \in C_i) = \frac{1}{2}(P(x \in C_i \mid S_1(x) = C_{j_{S_1}}) + P(x \in C_i) \mid S_2(x) = C_{j_{S_2}})) \tag{5.24}$$

Finally, the fused label will be determined by the next statement:

$$C_{fusion}(x) = \arg\max_{C_i} \{bel(x \in C_i)\} \quad i = 1, \cdots, N \tag{5.25}$$

Hence, the fused label will be computed by averaging the available RF probability vector. The problem of this approach is that all the classifiers have the same weight.

The following chapters present the experimental results for this work. Firstly, in Chapter 6, a evaluation of the single classifiers results is detailed. Secondly, the results of the fusion strategies presented and proposed in this chapter are shown. Lastly, a new estimation of the RF probability vector is proposed in this work and the results are given.

Part III

Experimental results

Chapter 6

Evaluating the prediction of the ensemble of classifiers composing the Random Forest

The main objective of this work is to present a fusion approach between classifiers in order to exploit the synergies between Sentinel-1 and -2. But, previously to this step and to support this goal, a set of analysis has been carried out.

This chapter presents the analysis of the RF predictions for the radar and optical classification chains. In order to accomplish with the analysis, the RF output has been modified to obtain the probabilities of belonging.

As seen in Section 3.3, Random Forest algorithm might be interpreted as an ensemble of *weak* classifiers. Each classifier is based on the technique of the binary decision-tree where each of them performs a decision for a given sample. In other words, in order to predict a given pixel, each tree "votes" for the class of belonging. Then, the class with a majority of votes is assigned as the output label. But this output only enable the user to know the assigned class for a given sample. For instance, consider a binary-class classification problem where two RF models are composed of ten trees ($K = 10$) each. Also, consider a new sample x, belonging to the class C_1, which falls in the trees of both models. The resulting probability vectors for this sample will be $p_{RF_1}(x) = \{10, 0\}$ and $p_{RF_2}(x) = \{6, 4\}$ for the first and second RF models, respectively. However the predicted label will be the same for both models, the exposed agreement between the decision trees is not the same. As it can be seen, the first case presents a total agreement which means that this model is 100% "sure" of the decision. Oppositely, the second model will release the same label but it shows an important level of uncertainty.

Typically, the most likely case is the second where not all the trees vote for the same class. So, as shown in the example, this classical approach implies considerable limitations since it is not possible to measure how "uncertain" is the model.

As explained in Section 3.3, the use of the probability vector allows to understand how the ensemble classifiers that built the forest are performing. For these reasons,

this chapter, and part of this work, is based on the use of the class probability vector as output of the RF algorithm instead of the standard label.

Firstly, this chapter presents a section where the probability vector is studied for the Sentinel-1 and Sentinel-2 classifiers providing an overview of them and comparing the obtained class probability distributions. Also, it is performed at class level in order to obtain an accurate analysis. Secondly, a set of visualization tools implemented to support the previous analysis are given.

6.1 Analysis of the Random Forest probability vector

As mentioned before, in a multi-source classification problem the class probabilities distribution could provide a significant amount of information. So, in order to get a proper understanding of this information the probability distributions can be related to each class prediction accuracy. Hence, given a class c_i, it can be interesting to study the class probability distribution categorizing them by three groups of samples:

(i) True Positive samples: this group represent those samples that have been correctly predicted (for a given pixel, the classifier predicts the class c_i and the reference data shows that belongs to class c_i).

(ii) False Positive samples: this group represent those samples that have been falsely predicted (for a given pixel, the classifier predicts the class c_i and the reference data shows that belongs to class c_j).

(iii) False Negative samples: this group represent those samples that have been erroneously predicted (for a given pixel, the classifier predicts the class c_j and the reference data shows that belongs to class c_i).

Consequently, the probability of belonging to each class, given a classified pixel, arises three different cases:

(i) The probability of belonging to the i-th class for a TP sample:

$$P(C_{pred.} = C_i \ / \ C_{ref.} = C_i)$$

(ii) The probability of belonging to the i-th class for a FP sample:

$$P(C_{pred.} = C_i \ / \ C_{ref.} \neq C_i)$$

(iii) The probability of belonging to the i-th class for a FN sample:

$$P(C_{pred.} \neq C_i \ / \ C_{ref.} = C_i)$$

Besides, two important parameters have been calculated for this work in order to exploit the extracted information. Details of each of them are given below.

The first parameter that it could be obtained from the class probability vector is a confidence metric which can be computed for each classified sample. In our multi-source classification problem this information could be very interesting since for each source, this metric allows to set its confidence level given a predicted sample.

Another parameter that can be extracted is the prediction margin. This margin is defined as the difference between the probability of the predicted class and the highest predicted probability of the other classes. The higher the margins is, the higher the agreement between the decision trees is. In other words, the margins display how well the classifier is discriminating between classes. Then, the margin parameter for a given sample can be defined as:

$$M(x) = \max_{c_i}\{p_{c_i}(x)\} - \max_{c_j}\{p_{c_j}(x)\} \qquad i = 1, \cdots, N \text{ and } j \neq i \tag{6.1}$$

where $p_{c_i}(x)$ is the probability of the predicted class and $p_{c_j}(x)$ the probabilities of the rest of classes.

Moreover, by means of the class probability distributions the histograms are presented in this chapter to allow a better interpretation of the given information. Accordingly, the margins histogram is also given to facilitate the data analysis.

As noticed before, the detailed analysis is carried out by using different sources in the classification system. Therefore, the RF outputs obtained from a single optical and radar classifiers are studied in this work.

6.1.1 Analyzing the Radar and Optical class probability results

As mentioned before, the analysis presented here is performed by classes in order to provide a better comprehension. So, the study of classes such as *Straw cereals*, *Vine* or *Orchard* are given.

For simplicity, not all the classes are shown. Due to the great amount of data only the more remarkable cases are going to be commented. All the studied cases might be found in the Appendix A.

For each class, we would like to study how the radar and optical classifiers performs. Therefore, in order to examine the set of predicted samples for each classifier, this work propose to visualize the results by means of three different figures. Firstly, probability and margins histograms are presented. Also, it is shown the direct correspondence between probabilities and margins. Finally, the last figure exhibited presents the direct relation between the probabilities for each single classifiers (*i.e.* radar and optical probabilities).

The classifications used for this study has been carried out using the same parameters for both chains. As shown in the configuration explained in Chapter 2 and 4 (see Table 4.1), the classification model has been trained with the following parameters:

Parameter	**Value**
Number of sets of pixels ($nbruns$)	10
Number of decision trees ($nbtrees$)	100
Number of training samples per class	2000
Number of dates ($nbdates$)	$nbdates_{rad.} = 60$
	$nbdates_{opt.} = 33$
Number of features per date ($nbands$)	$nbands_{rad.} = 2$
	$nbands_{opt.} = 10$
Number of total features (p)	$p_{rad.} = (60 \times 2)$
	$p_{opt.} = (33 \times 10)$
Max. number of features per tree (m)	$\sqrt{p}$
Max. depth of each tree (max_depth)	25
Min. sample leaf size ($min\_samples$)	25

Table 6.1: Training parameters for radar and optical classifications.

It is important to remark the fact that both classification results correspond to the 14th of October at the end of the agricultural season.

Straw cereal analysis

Histograms give an accurate representation of the distribution of numerical data. For these cases, it allows to interpret, in a graphical way, the distribution of the probabilities extracted from the set of decision trees.

Figure 6.1 presents a great example. Upper figure represents the distribution of the *Straw* probabilities for those pixels rightly predicted. Middle and bottom pictures show the probabilities distribution for those pixels erroneously classified corresponding to the False Positive and False Negative samples, respectively.

Comparing the three figures, it might be corroborated that most of the classified samples are well predicted. Also, these pixels have obtained high probabilities of belonging. By looking at TP samples figure, optical classifiers (*i.e.* green histogram) presents higher values than radar (*i.e.* blue histogram) meaning that optical samples are predicted with lower uncertainty.

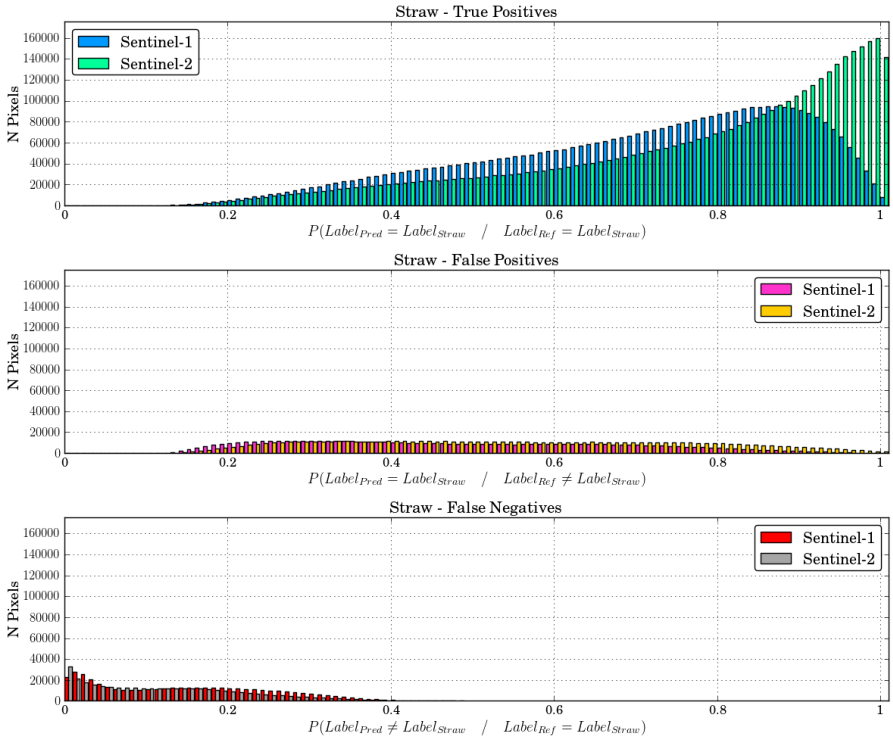


Figure 6.1: *Straw cereals* probability histograms for radar and optical classifications. Upper plot displays the True Positive probabilities distribution. Middle plot displays the False Positive probabilities distribution. Bottom plot displays the False Negative probabilities distribution.

Therefore, comparing the results obtained by the radar and optical classifications, it could be concluded that optical classifier is able to reach a greater accuracy for this class and besides, it presents an important confidence since the well predicted samples are related to higher probabilities (*i.e.* higher agreement between decision trees).

Concerning the FP and FN predicted samples, both classifiers presents a low concentration of these samples. Besides, one important remark comes from the fact that all the FN samples obtain low probability values meaning a strong confusion between the decision trees.

The second figure allowing to interpret the results is the margin histogram. As noticed before, the margin metric is another parameter to assess the robustness of the classifier measuring the agreement between the decision trees for a given sample. A high margin value means that there is a significant difference between the highest probability and the others, which implies that most of the trees have voted for the same class.

Figure 6.2 presents the margin histograms for the *Straw* class.

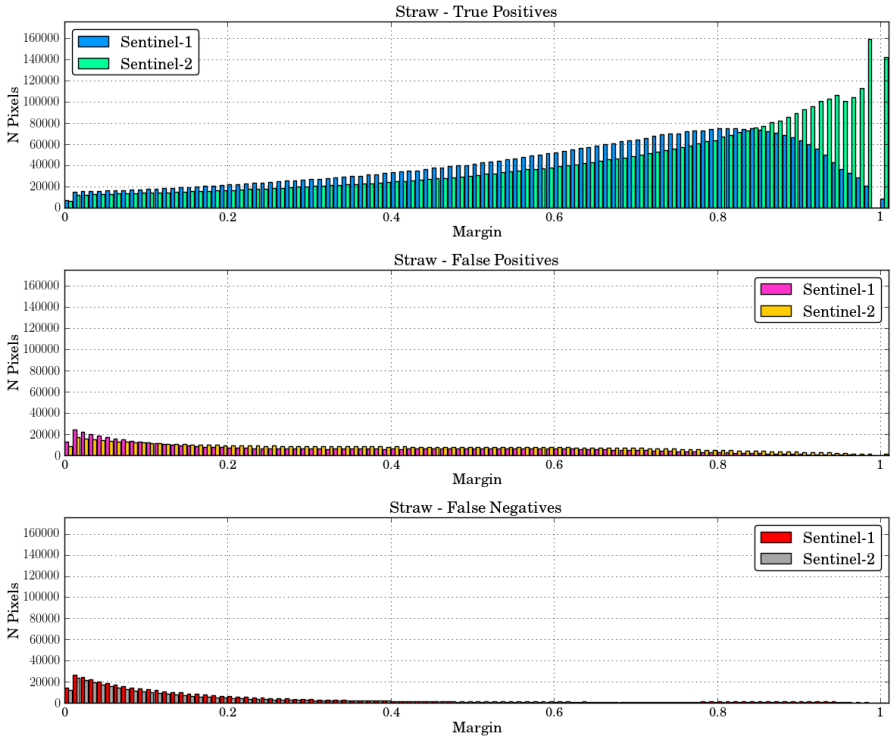


Figure 6.2: *Straw cereals* margins histograms for radar and optical classifications. Upper plot displays the True Positive margins distribution. Middle plot displays the False Positive margins distribution. Bottom plot displays the False Negative margins distribution.

Following the presented analysis structure, *Straw* margin histogram just confirms the formerly commented. Looking the figure related to the TP samples, optical classifier presents a higher number of samples at the highest margin values. This implies that optical model classifies with higher "confidence" than the radar model.

For this class, margins histograms show redundant information due to the limited uncertainty presented by both classifiers. But, it is worth commenting how the margin histograms for the FP and FN samples present low values. As mentioned before, low margin values involves a strong confusion for the decision trees of the classifiers. Also, if a sample is well predicted is desired a high margin value. Otherwise, it will be desired the lowest value possible, since if a sample is wrong predicted, a high margin value means that the model is erroneously classifying this pixel with an important "confidence".

Finally, the figure that allows to examine the relationship between the probabilities and the margins is presented. In other words, this figure allows to visualize the discernment between classifiers. Nontheless, Figure 6.3 presents this relation for the *Straw cereal* class without providing additional information since both classifiers show similar behaviours.

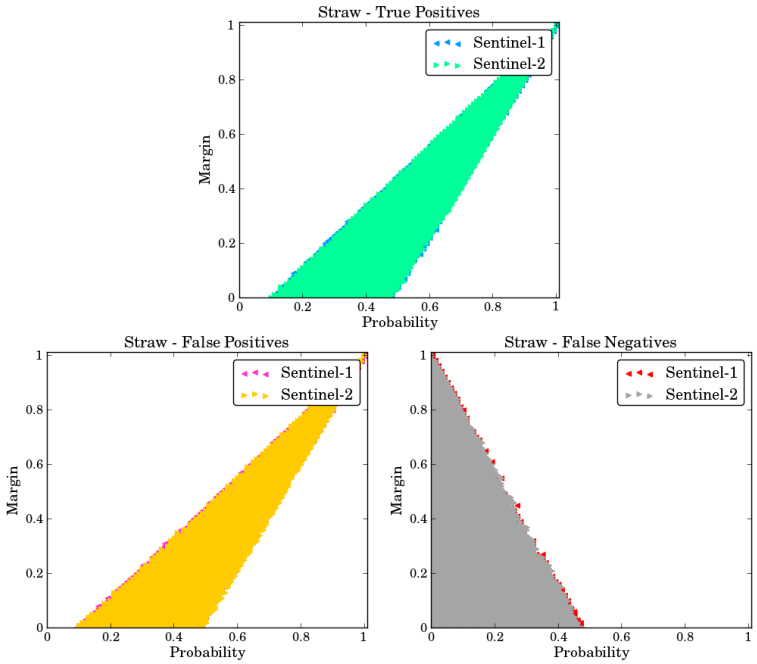


Figure 6.3: *Straw cereal* class probabilities vs. margins plots. Upper plot displays the TP samples. Bottom-left plot displays the FP samples. Bottom-right plot displays the FN samples. Each plot is presented for Sentinel-1 and Sentinel-2 classified samples.

Vine analysis

A second study is presented here for the *Vine* class. Figure 6.4 presents the probability histogram for the mentioned class.

Regarding the TP samples predicted by the radar classifier, Figure 6.4 shows how the histogram follows a normal distribution where the mean is located at the center of the figure (i.e. $\mu \approx 0.5$). Comparing this results with the ones provided by the optical classifier, it can be corroborated how the optical model is predicting the TP samples with higher probabilities. In fact, most of the samples are classified with a high probability unlike those values provided by the radar model.

Comparing with the resulting TP samples for *Straw cereal* class, radar classification presents a probability distribution where most of the pixels are well predicted too but they obtain lower probability values.

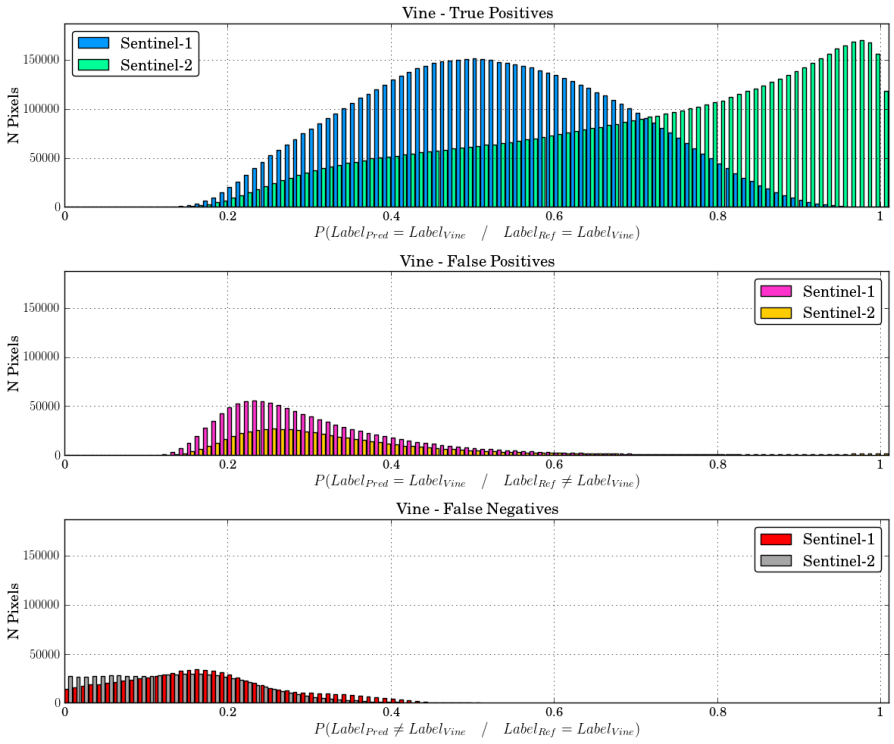


Figure 6.4: *Vine* probability histograms for radar and optical classifications. Upper plot displays the True Positive probabilities distribution. Middle plot displays the False Positive probabilities distribution. Bottom plot displays the False Negative probabilities distribution.

By looking at the FP samples figure, it is important to noticed how the radar classifier is performing since it presents roughly double samples wrong predicted. Despite this, both classifiers show low probability values for these samples. It is important to note how *Vine* class samples are gathered on low probability values while *Straw cereal* samples present a larger range of values. This means that radar and optical models are misclassifying these pixels with more "confidence" for the *Straw cereal* case.

Following the structure of the analysis, Figure 6.5 presents the margin histograms for the *Vine* class.

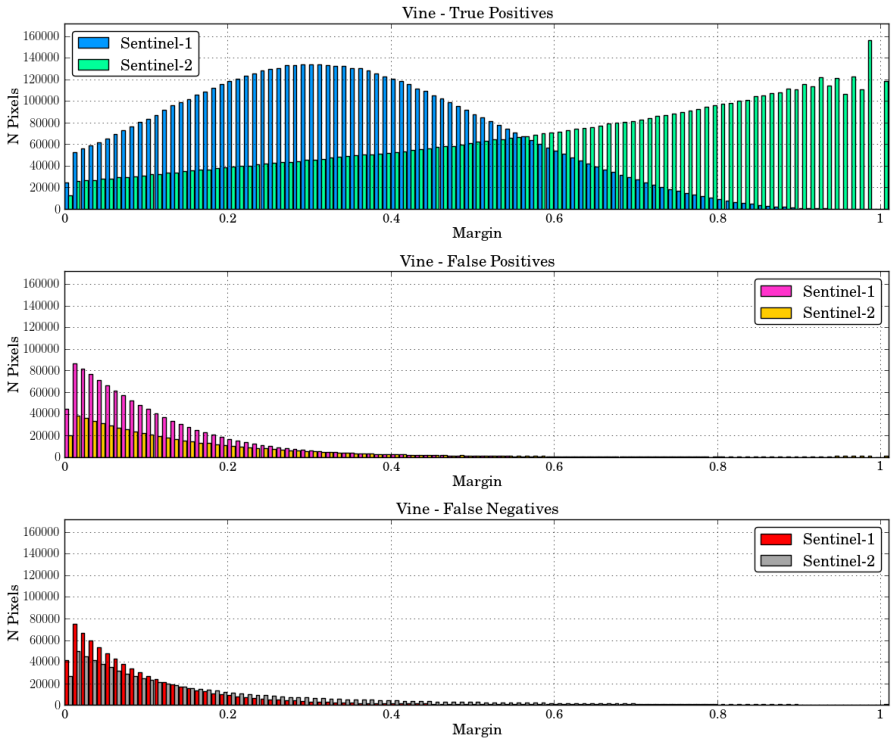


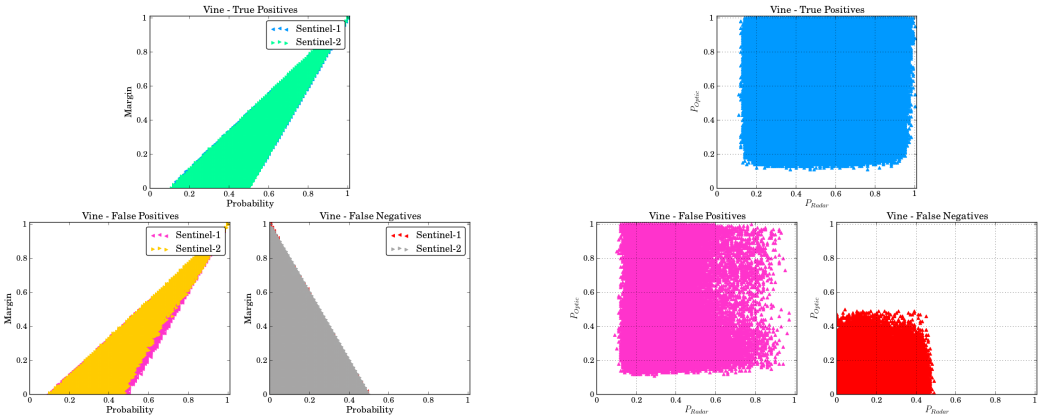
Figure 6.5: *Vine* margins histograms for radar and optical classifications. Upper plot displays the True Positive margins distribution. Middle plot displays the False Positive margins distribution. Bottom plot displays the False Negative margins distribution.

As mentioned before in Section 3.3, the class probabilities are a measurement about how well the RF trees are voting. So, if a pixel present a high probability for a given class, means that there is a broad agreement among the decision trees. Consequently, and regarding TP and FP samples, the lower the probability value is, the lower the margin value is. This reasoning could be corroborated by Eq. 6.1, since there is a linear relationship between the margin parameter and the probability for a given class.

Following this reasoning, the TP samples shown in Figure 6.4, express a limited agreement for the radar classifier. In the same way, upper plot from Figure 6.5 allows to illustrate the explained before. Besides, regarding the radar classifier, margin values show a normal distribution with a wide deviation and centered on a low probability value. This means that even though these samples are well predicted the radar classifier "hesitates" more than the optical.

Finally, low margin values shown by FP and FN samples in Figure 6.5 demonstrates that the confidence obtained by the pixels that have been missclassified is low for both classifiers.

Figure 6.6a just corroborates the reasoning explained above. For the samples predicted as the given class, it exists a linear correlation between the probability and the margin values.



(a) (b)

Figure 6.6: (a) Probabilities vs. Margins for the *Vine* class. (b) Radar vs. Optical probabilities for the *Vine* class. For Figure (a) and (b) upper plots display the TP samples. Bottom-left plots display the FP samples. Bottom-right plots display the FN samples.

It is important to remark what Figure 6.6b illustrates. This figure shows, for a given pixel, the probability value obtained by the optical and radar classifier. Hence, it is able to check how the models are behaving for the same sample. By looking on the TP and FP samples figures, it is possible to see how optical classifier is given higher probability values. This fact has a twofold interpretation. On hand hand, optical classifiers is more confident when predicts well than radar. But on the other hand, it is also more confident when misclassifies.

Orchard analysis

Finally, the same study is carried out by the *Orchard* class. Firstly, the probability histogram is presented. As the Figure 6.7 shows, this class obtains a poor classification performance.

Comparing the three figures, it is evident that both classifiers are not able to achieve a large number of well predicted samples. Concerning the TP results is important to remark the low probability values obtained. Despite this, the optical classifier presents a large histogram deviation what implies that it is achieving higher probability values than radar.

FP probability histogram shows how most of the classified pixels are falsely predicted as *Orchard* class. Another remark can be observed by comparing the classifiers results is the major presence of those samples predicted by the radar classifier.

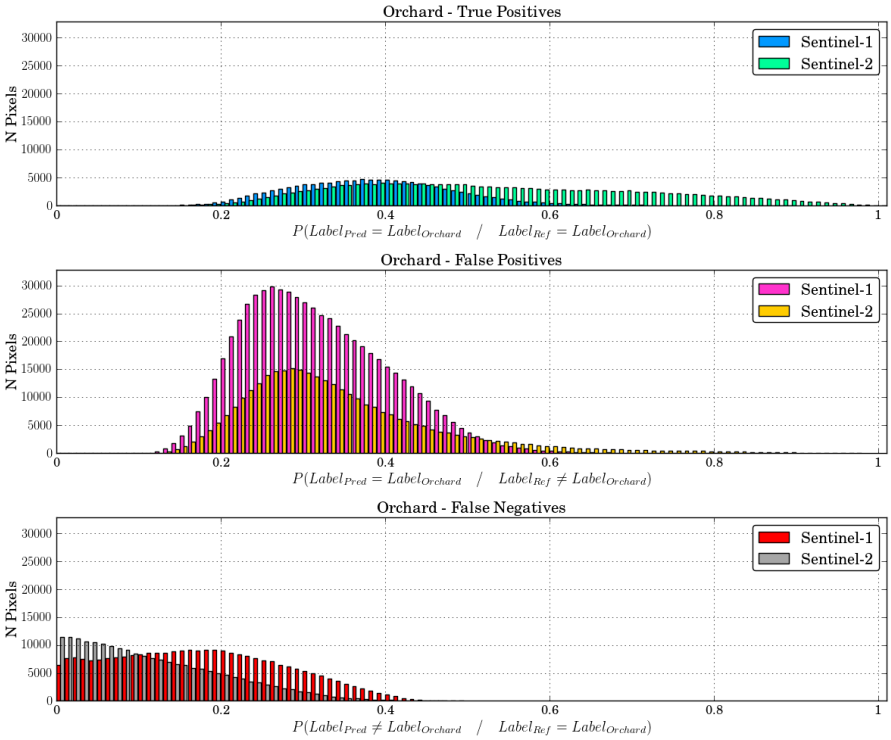


Figure 6.7: *Orchard* probability histograms for radar and optical classifications. Upper plot displays the True Positive probabilities distribution. Middle plot displays the False Positive probabilities distribution. Bottom plot displays the False Negative probabilities distribution.

Besides, another interesting observation is the fact that TP, FP and FN samples show similar probability ranges. By looking on the three figures, it exists a great amount of predicted samples that obtain a probability value between 0.2 and 0.4. This implies that the classifiers does not relate the accuracy class with the probability value.

Figure 6.8 plots the histograms of the obtained margin values for the *Orchard* class. As it can be seen in the previous class studies, the margin histogram shows similar information than the histogram of the probability values. Hence, looking at these figures it can be noticed that an important number of FP samples are predicted by both classifiers. Also, radar classifier predicts a higher amount of FP samples with a low margin value.

For those samples that are missclassified it can be understood as the classifier is "mistaken" or "being confused". Then, it can be expressed as "confusion" samples. Confusions between classes will be discussed further in this work in order to analize how and why the classifiers are wrong.

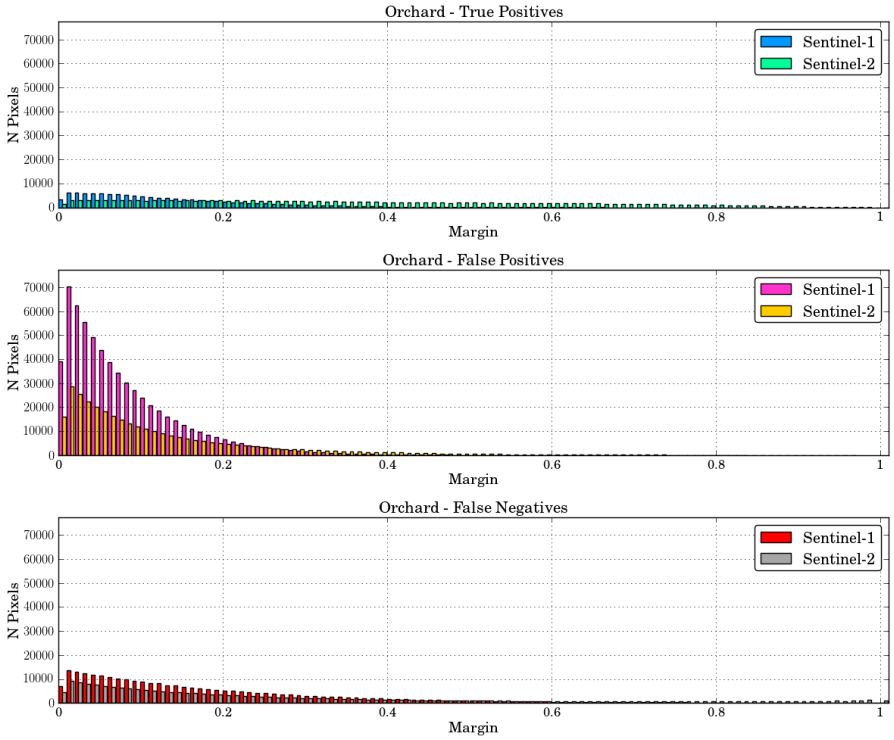


Figure 6.8: *Vine* margins histograms for radar and optical classifications. Upper plot displays the True Positive margins distribution. Middle plot displays the False Positive margins distribution. Bottom plot displays the False Negative margins distribution.

Finally, Figure 6.9 is presented. These three figures show the predicted pixels distributed according to the probability value assigned by the radar and the optical classifier. As explained before, it helps to see how performs both classifiers given a certain pixel. An interesting remark about the TP and FP samples is the distribution that these pixels show. As seen in the two previous class analysis, a pixel predicted by the optical classifier obtains higher probability value than by the radar.

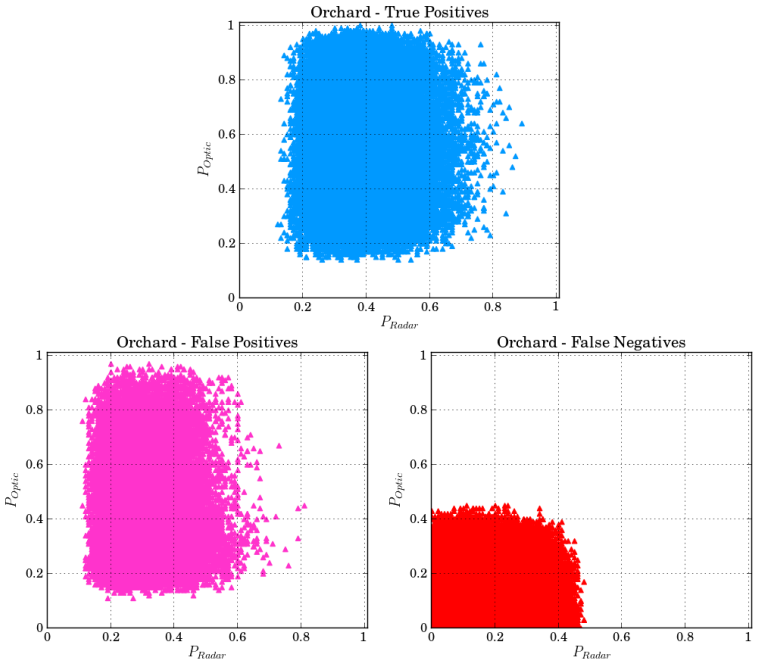


Figure 6.9: Radar vs. Optical probabilities for the *Orchard* class. Upper plot displays the TP samples. Bottom-left plot displays the FP samples. Bottom-right plot displays the FN samples.

This section has analyzed different classification results obtained by different classes. The studies has been based on the RF probabilities obtained for each classification. This extracted information has been exploited in different ways. It has shown that each classifier obtains different "confidence" level for the same samples. Accordingly, the use of RF probabilities, instead of labels, implies a better knowledge of the performance. Therefore, in order to define a fusion classification approach a deeper knowledge could imply a better accuracy.

6.2 A visual evaluation of radar and optical classification results

A primary goal of data visualization is to communicate information clearly and efficiently. Hence, it makes complex data more understandable and usable. Thus, in order to complement the statistical analysis in Section 6.1, this section proposes a visual evaluation of the radar and optical classifications by computing a set of classification maps.

The classifications results used for the purposes of this section has been obtained by means of the same trained models than Section 6.1 (For more details about the training parameters see Table 6.1).

Classification maps are built by applying the trained classifier model on all the satellite image pixels. Therefore, these maps will show different results depending on the output of the classifier. As explained before, the classifier output could consist in a class label or a class probability vector. Thus, in this work two different maps are presented. Also, these maps are shown along the agricultural year. The goal is to shown how the maps may change along the time and therefore, how the classifier model is changing.

Firstly, the classification maps are shown. These map are built assigning to each pixel the predicted label.

Secondly, the confidence maps are presented too. As commented before, these maps can be built by means of the predicted labels or the class probabilities. Thus, the confidence maps are built assigning the highest probability to the classified pixel. These maps are an important evaluation tool for the user since it shows how confidence is the classifier model.

Classification maps

In order to illustrate the characteristics of these maps, each map is presented for three different dates along the agricultural season. And for the sake of a correct comparison, the classification maps are displayed for radar and optical classifiers.

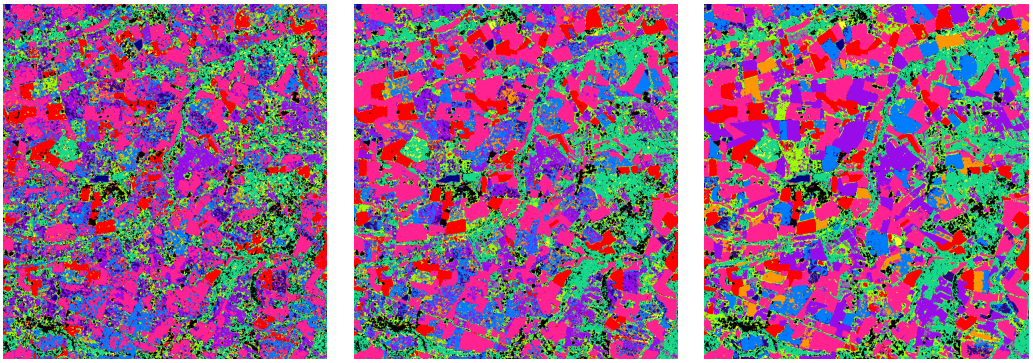
Figure 6.10a represents an small zone of the study area which represents a region about $20km^2$ around the city of Toulouse (France). This image has been acquired from the Sentinel-2 satellite and it illustrates a RGB color composition.

(a) (b)

Figure 6.10: (a) RGB image from Sentinel-2 presenting the study area (b) legend for the classification maps

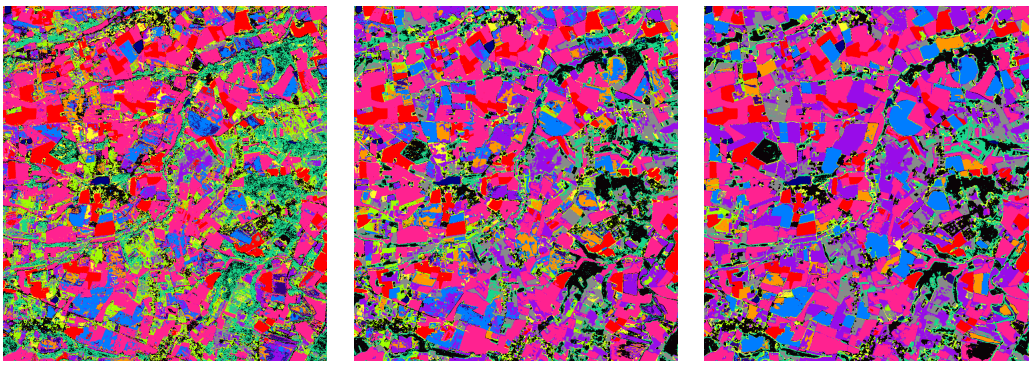
In Figure 6.11 three classification map are shown which have been obtained by means of the radar classification models for three different dates along the season. In contrast, Figure 6.12 presents the corresponding classification maps obtained with the optical model.

Classification maps for the 8-9th of October date show how the classification models present a confusion between *Shrubland* and *Build up* classes. This can be observed mostly on the right part of the maps. It exists an important number of areas where pixels predicted as *Shrubland* and *Build up* classes behave as impulse noise. Comparing radar and optical results, it may be seen how the radar classification model tends to increase the number of predicted pixels as *Shrubland* class. This trend can be noticed by comparing the classification maps for the different dates. In contrast, the optical results, presents the same trend by predicting pixels as *Build up* class. Therefore, as season goes by each classifier presents a different trend for those pixels. Then, it may be said that each classifier tries to solve these "confusions" in a different way.



(a) 9th December (b) 29th April (c) 14th October

Figure 6.11: Map composed by the class labels obtained from the Radar classification chain and for three different dates concerning the beginnings, mid and the end of the agricultural season



(a) 8th December (b) 27th April (c) 14th October

Figure 6.12: Map composed by the class labels obtained from the Optical classification chain and for three different dates concerning the beginnings, mid and the end of the agricultural season

Another interesting remark is how are predicted the pixels belonging to the Soybean class. Crops classes can be divided as winter or summer crops, depending on the weather they need for best growth. Soybean is a species of legume that requires warm soil and high temperatures to grow therefore it is classified as a summer crop.

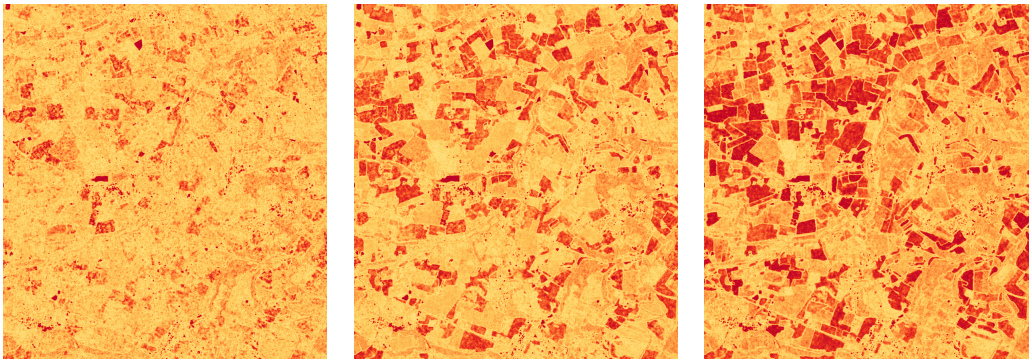
Consequently, by looking on the classification maps it can be noticed how *Soybean* pixels (*i.e.* orange pixels) only appears after spring.

As may be seen, the labeled maps are not helpful to evaluate the classification accuracy but it could be a powerful visualization tool in order to analyze the classification evolution along the seasons and to compare the labeled pixels for different sources.

Confidence maps

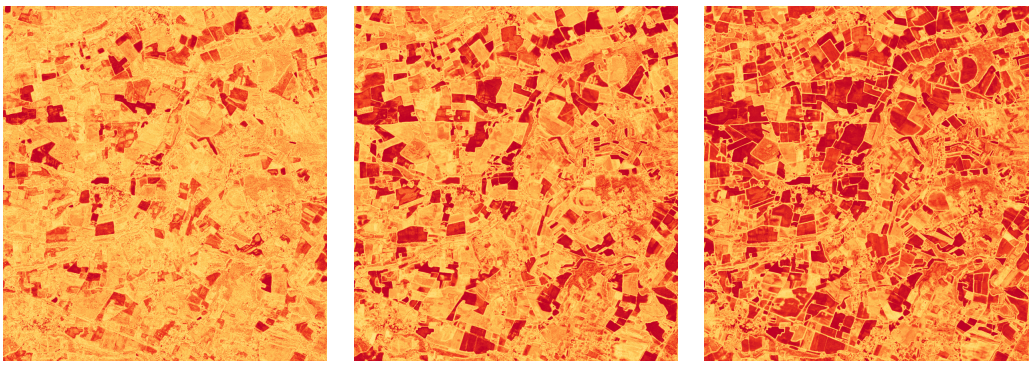
As commented before, the confidence maps are a set of maps where the value of each pixels corresponds to the value of the highest probability obtained by the RF model (See Section 3.3). By plotting this value on a map, it is possible to check the "conviction" of the classifier.

In Figure 6.13 and Figure 6.14 the confidence maps for the radar and optical classifications are shown. These maps are presented for three different dates along the agricultural season.



(a) 9th December (b) 29th April (c) 14th October

Figure 6.13: Confidence maps obtained from the radar classification results and for three dates concerning the beginnings, mid and the end of the agricultural season.



(a) 8th December (b) 27th April (c) 14th October

Figure 6.14: Confidence maps obtained from the optical classification results and for three dates concerning the beginnings, mid and the end of the agricultural season.

For these images, the confidence level is represented by using a red scale color. The highest confidence level in the scale is represented with a saturated color red. As the confidence level decreases, corresponding color tones become closer to the yellow color.

Comparing Figure 6.13 and Figure 6.14, it can be observed how the confidence maps obtained from optical results present more reliable predictions since red levels are higher. Besides, for both cases it can be noticed how the pixels are getting more red as season goes by. This means that the classification models are learning and improving the conviction on its predictions.

Therefore, these maps shows that the optical classification obtains higher probability values for the predicted samples and accordingly it predicts with higher "confidence". Nonetheless, it should be mentioned that the achievement of high probability values does not imply a correct prediction as seen in Section 6.1.

6.3 Conclusions

Different studies have been carried out here in order to evaluate the predictions of the ensemble of weak classifiers composing the Random Forest algorithm. The main goal has consists in to study the class probability vector that can be obtained as an output classification result instead of the class label. In this study, the information contained on the probability vector has been studied in the context of the radar and optical classification. The interest is to study the information contained in these vectors when the input data is not the same.

Firstly, the information contained in the class probability vector has been analyzed in different ways. For this purpose, a set of statistical figures based on the radar and optical classification results has been shown. From this first study, it could be seen how there are classes (*e.g. Straw cereal* class) easier to classify by the RF model. These classes involve high probabilities and low margins values. Instead, other classes such as the *Orchard* class, implies important difficulties to be correctly classified. Besides, it has been shown how different input data, carries different classification performances. Some differences have been observed by comparing optical and radar results. Here, it was shown how for a given pixel, radar and optical classifiers obtain different probability values where the higher value belongs to the optical model for most of the cases.

Finally, a visual evaluation has been carried out for radar and optical classification results. Two different maps has been shown. Firstly, it was presented the classification map where all the pixels correspond to the predicted labels. Secondly, the confidence map was shown. These maps are built assigning the highest probability value to the predicted pixel. This visual evaluation has corroborated the fact that optical classifier predicts with a higher "confidence" level.

Chapter 7

Evaluation of the fusion classification strategies

Multi-source data fusion for land cover purposes has become into a hot-topic for many researches due to the necessity to exploit the features from different Remote Sensing sources. As presented in Section 1.3, this work aims at studying different classifier fusion techniques. The main goal is to develop a framework able to combine the optical and SAR sensor data from Sentinel-1 and -2, respectively. Specifically, Chapter 5 presented five different fusion methods at decision level which has been implemented and tested. Besides, the interest of this chapter is to show how fusion techniques exploits the information provided by the class probability vector being able to achieve better accuracies.

For the purposes of this chapter, radar and optical classification results are used. Hence, in order to guarantee the consistency of the results the classification framework is the same as defined in Section 6.1. Table 6.1 shows the details of the configuration parameters for each classification chain.

Fusion strategies presented in this work will be compared with single classifiers (*i.e.* radar and optical classifications) but also with the most classical approach that can be performed in order to combine optical and radar data. This approach consist in using all the available optical and radar remote sensing data ensemble as input data in the classification system. As commented in Section 5.1, this approach is known as a combination step at pixel level. Therefore, this chapter aims to study how fusion at decision level is able to improve the accuracy for a single classifier but also for the fusion approach at pixel level.

The following chapter presents the results for the five combination techniques studied in this work. Firstly, the results obtained by the fusion strategies are presented in Section 7.1. These results are obtained by using the evaluation metrics detailed in Section 3.4. The evaluation will be presented by considering the temporal dimension of the input data. Therefore, the different metrics will be computed at different instants of time. Secondly, in Section 7.2 the evaluation presented studies the wrong predictions by studying the confusions obtained. Besides, it is also

studied the agreement between classification approaches.

7.1 Statistical temporal evaluation

This section shows the results of the fusion approaches presented in Chapter 5. The goal is to compare the results obtained by the five fusion techniques and the results obtained by the single classifiers presented in Section 6.1 and the combination approach at pixel-level. Then, this study is performed by comparing eight different classification strategies. Below is described each classification approach:

(i) Sentinel-1 (S1) classification. Classification characterized by the use of radar data as input data.

(ii) Sentinel-2 (S2) classification. Classification characterized by the use of optical data as input data.

(iii) Sentinel-1 plus Sentinel-2 (S1S2) classification. Classification characterized by the use of all available remote sensing data. Radar and optical data is used ensemble as input data. The data is merge at pixel level.

(iv) Dempster-Shafer (DS) classification. Classification characterized by a fusion stage at decision level. This fusion technique merges the resulting data from S1 and S2 classifications by means of the so-called Dempster-Shafer Theory (See Section 5.2 for more details).

(v) Bayes Belief Integration (BB) classification. Classification characterized by the a fusion stage at decision level. This fusion technique merges the resulting data from S1 and S2 classifications by means of the Bayes Belief Integration method (See Section 5.3 for more details).

(vi) Modified Dempster-Shafer (M-DS) classification. Classification characterized by a fusion stage at decision level. This fusion technique merges the resulting data from S1 and S2 classifications by means of the Modified Dempster-Shafer method (See Section 5.5 for more details).

(vii) Median Rule (MR) classification. Classification characterized by the a fusion stage at decision level. This fusion technique merges the resulting data from S1 and S2 classifications by means of the Median Rule method (See Section 5.6 for more details).

(viii) Maximum Confidence (MC) classification. Classification characterized by the a fusion stage at decision level. This fusion technique merges the resulting data from S1 and S2 classifications by means of the Maximum Confidence method (See Section 5.4 for more details).

The proposed fusion strategies may be divided in two sub-groups depending on the RF output used. On one hand, the *true probabilistic methods* are those methods

that the merge stage is purely based on the use of the class probability vectors. BB, MR and MC classifications are the approaches belonging to this category. On the other hand, the *Dempster-Shafer methods* are based on the mathematical Evidence Theory (explained in Section 5.2). These methods use the confusion matrices and the predicted labels to perform the merge of the classifiers decisions. DS and M-DS approaches belongs to this sub-group.

For all the eight presented approaches, different metric are evaluated in the following section. As explained before, metrics are computed at different times in order to obtain a temporal evaluation. Concerning the evaluation metrics four classical metrics are shown. Firstly, the Overall Accuracy metric is given in order to obtain a general overview of how are performing each of the eight approaches. Secondly, in order to show how the different classifications are performing by classes the Precision, Recall and F-Score metrics are presented. Besides, all the evaluation metrics are shown in % an averaged over the total number of sets of pixels (*i.e. nbruns*).

Accordingly with the report structure, for simplicity, not all the classes are shown. Due to the great amount of data, only the most interesting cases are commented. For more details see Appendix B.

7.1.1 Overall Accuracy

As explained in Section 3.4, the Overall Accuracy (OA) was defined as a metric to show the probability that a sample is correctly classified. Using the Equation 3.5, the OA values computed for the eight classification approaches on different dates are plotted in Figure 7.1. As it may be seen, S1 classification obtains the worst results. But, despite this, this approach obtains the highest improvement along the time. Following the results descriptions, it is important to remark the time evolution of the S2 and S1S2 classifications. Both approaches obtains similar results at the end of the season being the S2 classification the best of both. But, it is interesting to observe that at the beginnings of the season the S1S2 classification obtains better results. This phenomenon can be explained by the high-dimensional future space characterizing the S1S2 approach. This phenomenon is explained in Section 3.1 and it is known as the *curse of dimensionality*.

Regarding the proposed fusion strategies, the results show that the fusion step may improve the previous results. Besides, the different approaches follow manly two different behaviours. On one hand, the BB and MR fusion strategies achieve the best results being the BB classification the best one. On the other hand, the DS, M-DS and MC fusion strategies obtain better results than the S2 and S1S2 classifications but worse than BB and MR approaches. As explained before, fusion strategies may be divided in two categories. Therefore, with the exception of the MC fusion, true probabilistic methods are able to obtain the best OA results. But, BB and MR fusion techniques not only reach the best OA at the end of the season, they also outperform the other strategies all along the agricultural season.

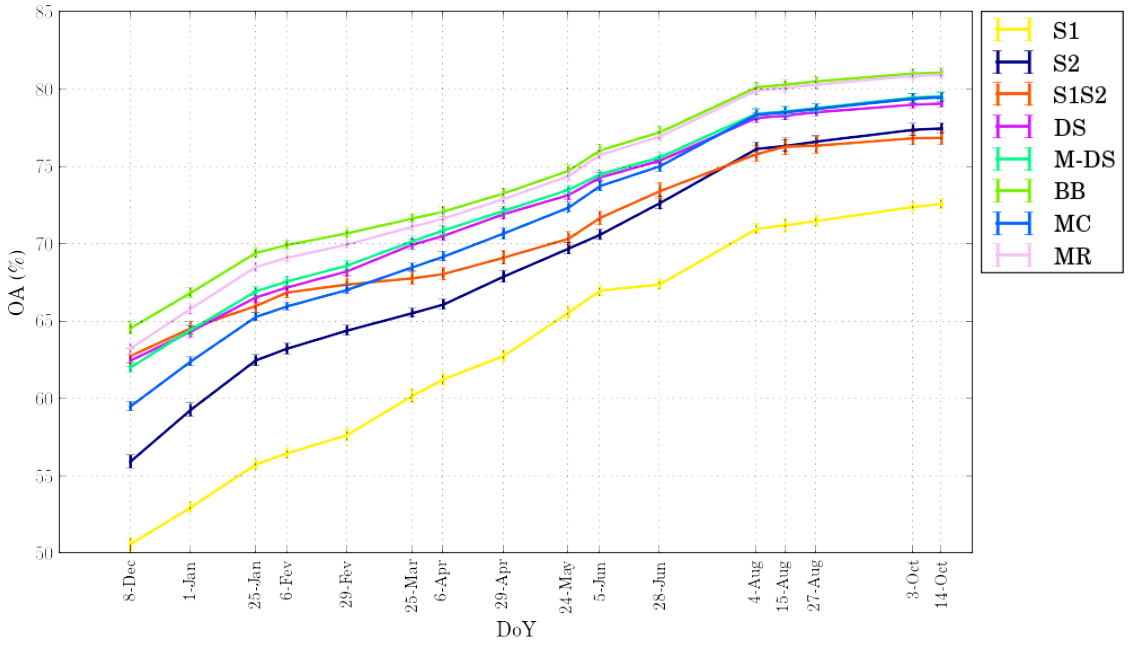


Figure 7.1: OA metric along the season for the five presented fusion techniques (DS, BB, M-DS, MR and MC), and the S1, S2 and S1S2 classifications. The metric is given in % and it is averaged over 10 runs.

In Table 7.1 it can be seen the differences in % for the OA results between the S1, S2 and S1S2 classifications and the presented fusion strategies. This table shows in a different way the fact that even the worst fusion performance is able to obtain higher results.

DS			M-DS			BB			MC			MR		
S1	S2	S1S2	S1	S2	S1S2	S1	S2	S1S2	S1	S2	S1S2	S1	S2	S1S2
6.5	1.6	2.2	7.0	2.1	2.7	**8.5**	**3.6**	**4.2**	6.9	2.0	2.6	8.4	3.5	4.1

Table 7.1: OA improvement in % averaged over 10 runs for the 14th of October.

Nonetheless, as mentioned in Section 3.4, global metrics such as OA are not enough to evaluate the quality of the classification performance. Consequently, in order to perform a right assessment, metrics by class are presented in the following subsection.

7.1.2 Precision, Recall and F-Score

First of all, an overview of the results obtained by all the proposed classification strategies is given by Table 7.2. This table shows the F-Score results for the 14th October which corresponds to the end of the agricultural season. The presented results are shown for the S1, S2 and S1S2 classifications and the proposed fusion strategies.

By looking at Table 7.2, it may be observed how the BB fusion method outperforms the rest of classification strategies for most of the classes. Therefore, these results corroborates the results show in Figure 7.1. Despite of the great results of the BB approach, it can be remarked that some classes such as Orchard, Grassland, Soybean or Sorghum reach better classification results for other strategies.

	14 Oct							
Class	S1	S2	S1S2	DS	M-DS	BB	MC	MR
Vine	87.4	90.8	91.2	89.1	89.7	**94.3**	92.9	94.0
Straw	88.5	88.2	89.1	88.0	88.4	**90.4**	89.5	90.1
Maize	83.0	85.4	84.1	86.9	87.0	**88.2**	87.1	88.1
Sorghum	31.3	40.8	35.8	**48.3**	47.7	45.1	42.6	44.9
Soybean	58.4	58.8	59.3	64.3	64.1	67.6	65.7	**67.8**
Sunflower	86.7	82.9	85.9	85.5	85.1	**87.1**	86.0	86.7
Alfalfa	53.1	57.9	57.7	55.1	57.1	**64.5**	61.5	64.1
Grassland	62.0	72.2	69.2	75.4	**75.8**	73.6	72.1	73.7
Fallow	41.3	49.6	49.0	44.2	46.7	**54.5**	51.7	54.3
Shrubland	49.0	48.3	52.0	48.3	48.8	**58.1**	54.7	58.0
Raspseed	77.5	71.9	75.0	75.8	75.5	**77.6**	76.3	77.3
Deciduous	81.7	85.4	86.8	83.9	84.0	**87.8**	86.0	87.3
Evergreen	33.3	76.0	72.3	76.1	75.2	**81.1**	64.8	74.2
Build up	87.2	89.5	93.2	90.4	91.2	**96.1**	93.2	95.0
Water	95.6	**98.8**	98.3	97.3	97.3	98.2	98.2	98.4
Orchard	20.2	40.8	32.3	31.0	35.5	43.6	41.4	**45.4**
OA	72.6	77.5	76.9	79.1	79.6	**81.1**	79.5	81.0
95% Confidence	0.49	0.53	0.64	0.37	0.36	0.39	0.46	0.41

Table 7.2: F-Score in % averaged over 10 runs.

Moreover, the Confidence Interval (CI) is presented at the bottom of the table. As explained in Section 3.4, the CI provides a range of values within which the population parameter is likely to lie. Therefore as may be seen in Table 7.2, the proposed fusion strategies obtain lower values. The lowest values are obtained by the Dempster-Shafer methods and the S1S2 classification present the highest range.

It is worth mentioning that the fusion techniques based on the handling of the probability vectors achieve significant similar results. Also, it is important to remark the interesting cases of the *Sorghum* and *Grassland* classes which achieve greater results with the methods based on the Dempster-Shafer theory.

Besides, Figure 7.3 shows, in a similar way than Figure 7.1, the improvement for the F-Score results (in %) between the S1, S2 and S1S2 classifications and the BB fusion. It is interesting to remark how the BB is able to outperform for all the classes with the exception of the *Water* class.

14 Oct	Bayesian Belief Integration		
Class	S1	S2	S1S2
Vine	6.9	3.5	3.1
Straw	1.9	2.2	1.3
Maize	5.2	2.8	4.1
Sorghum	13.8	4.3	9.3
Soybean	9.2	8.8	8.3
Sunflower	0.4	4.2	1.2
Alfalfa	11.4	6.6	6.8
Grassland	11.6	1.4	4.4
Fallow	13.2	4.9	5.5
Shrubland	9.1	9.8	6.1
Raspseed	0.1	5.7	2.6
Deciduous	6.1	2.4	1.0
Evergreen	47.8	5.1	8.8
Build up	8.9	6.6	2.9
Water	2.6	-0.6	-0.1
Orchard	23.4	2.8	11.3
OA	8.5	3.6	4.2
95% Confidence	-0.1	-0.1	-0.2

Table 7.3: F-Score improvement in % averaged over 10 runs.

Following, a temporal evaluation of the Precision, Recall and F-Score metrics by certain classes is presented (See Appendix B for more details).

Vine class metrics

The interest of studying the *Vine* class is explained by the fact that this class obtains remarkable results for S1 and S2 classifications. Besides, as Section 6.1.1 shows, *Vine* class presents very different probability histograms despite they are able to correctly predict an important amount of pixels.

Figure 7.2 presents the Precision, Recall and F-score metrics for the *Vine* class. As it may be seen, upper figures correspond to the Precision and Recall metrics, respectively on the left and right sides. Finally, the F-Score metric corresponds to the left-bottom figure.

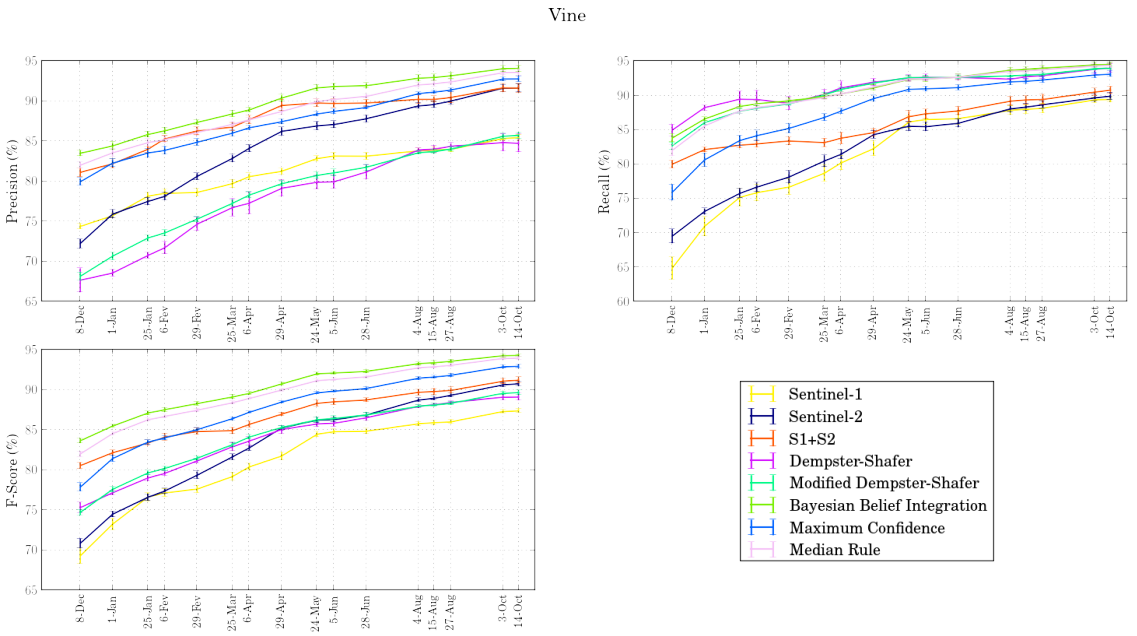


Figure 7.2: Temporal evaluation metrics. Precision, Recall and F-Score results along the agricultural season for the *Vine* class.

Vine class meets an early success for all the metrics. Regarding the Precision plot, only the fusion techniques purely based on the class probabilities achieve better rates than S2 and S1S2 classifiers. Also, an interesting observation is how the DS methods are biased by the Radar tendency achieving the worst results.

In contrast, DS methods shows a great performance for the Recall metric. This fact implies that DS methods are able to reduce an important number of FN samples from S1 and S2 results despite the cost is a poor precision metric. Also, the Recall plot shows how the proposed fusion strategies outperform to the S1, S2 and S1S2 classifications.

Finally, looking at the F-Score figure it may be seen that the BB approach obtains the best results along time and is followed by the MR and MC fusions. Consequently, these results show how the true probabilistic methods are able to remove a great amount of FP and FN samples. In contrast, DS and M-DS methods are only able to outperform the S1 classifier therefore it obtains worse results than S2 and S1S2 classifications.

Build up class metrics

Another interesting class to study in this section is the *Build up* class. It is interesting since it is considered a permanent class such as *Water* class. This means that these classes do not change over time, therefore they present the same characteristics along the season. Besides, these classes are able to obtain high probability values along the temporal dimension. Hence, pixels belonging to these classes are classified with a high "confidence" level.

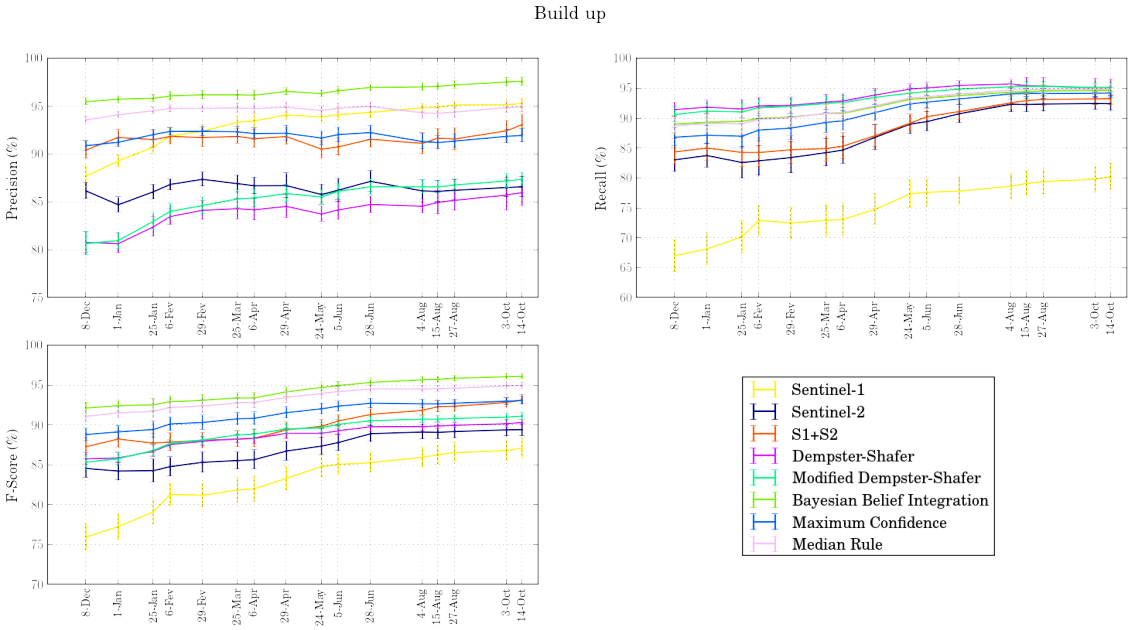


Figure 7.3: Temporal evaluation metrics. Precision, Recall and F-Score results along the agricultural season for the *Build up* class.

Figure 7.3 show the Precision, Recall and F-Score results for the *Build up* class. This is an interesting class since it obtains, together with the *Water* class, the best results for all the strategies. As commented before, this class does not change over time and this fact it can be seen in the results.

As it may be seen in most of the class evaluations, DS methods presents a particular behaviour. The higher the precision values are, the lower the recall values, and vice versa. In other words, if the number of FN samples are reduced, the number of FP samples increase. For the studied class, DS methods present the worst precision values and the best recall results. Consequently, they only are able to outperform the S1 and S2 classifiers for the F-Score results.

An interesting observation is how the S1 classification is only outperformed by the BB fusion for the precision results. And, despite this good results, it obtains the worst recall values.

Finally, true probabilistic methods obtain the best results for the F-Score values. Specially, the BB fusion is able to achieve the best precision and F-Score values.

Sunflower class metrics

Here, it is detailed the temporal evaluation of the *Sunflower* class. As in former cases, Figure 7.4 presents the Precision, Recall and F-Score results. The *Sunflower* class is an interesting case since it is considered as a summer crop which means that it starts the growth after the spring. By looking at Figure 7.4, it can be observed how there is an important improvement of the results as summer period approaches. Thus, for summer classes, the temporal evaluation is particularly important.

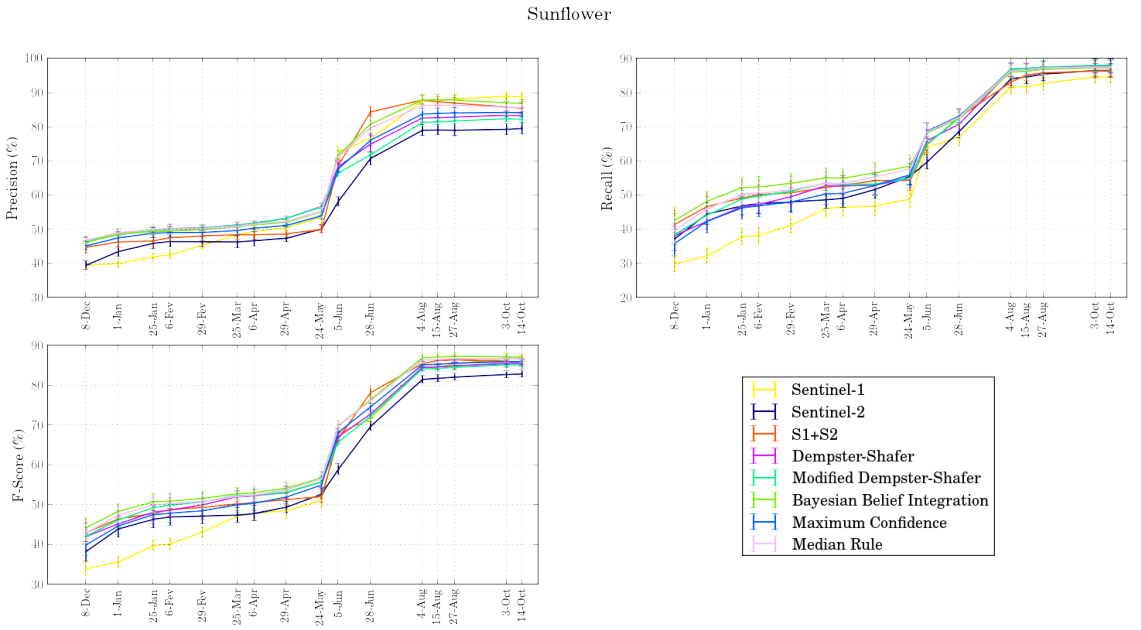


Figure 7.4: Temporal evaluation metrics. Precision, Recall and F-Score results along the agricultural season for the *Sunflower* class.

Concerning the metric results shown by Figure 7.4, the proposed fusion strategies do not involve a significant improvement. In fact, all the proposed strategies obtains similar results. Despite that, regarding the F-Score figure, BB fusion is able to achieve the best results.

Besides, it is important to remark that, the S1 classification is only outperformed by the BB strategy for the F-Score results and it presents the best precision values. This is an interesting observation since the S1 classifier seldom outperform the S2 classifier.

Evergreen class metrics

Figure 7.5 presents the temporal evaluation for the *Evergreen* class. This class obtains significant different results for S1 and S2 classifiers. In the case of the S2 classification results, it can be observed at Figure 7.5 how the high performance values are obtained along all the agricultural year. In contrast, the S1 classifier shows the worst results for the three metrics. The S1 classification presents a high number of FP samples for the *Evergreen* class. This is due to the fact that the S1 model is confusing between *Evergreen* and the other permanent classes. Besides, at the beginnings of the season it shows low recall values. This can be explained since the S1 model shows a high number of FN samples at this instant of time. Analyzing these misclassified samples, the S1 classifier shows an important confusion between *Evergreen* and *Deciduous* classes until the summer period comes (See Appendix B for more details).

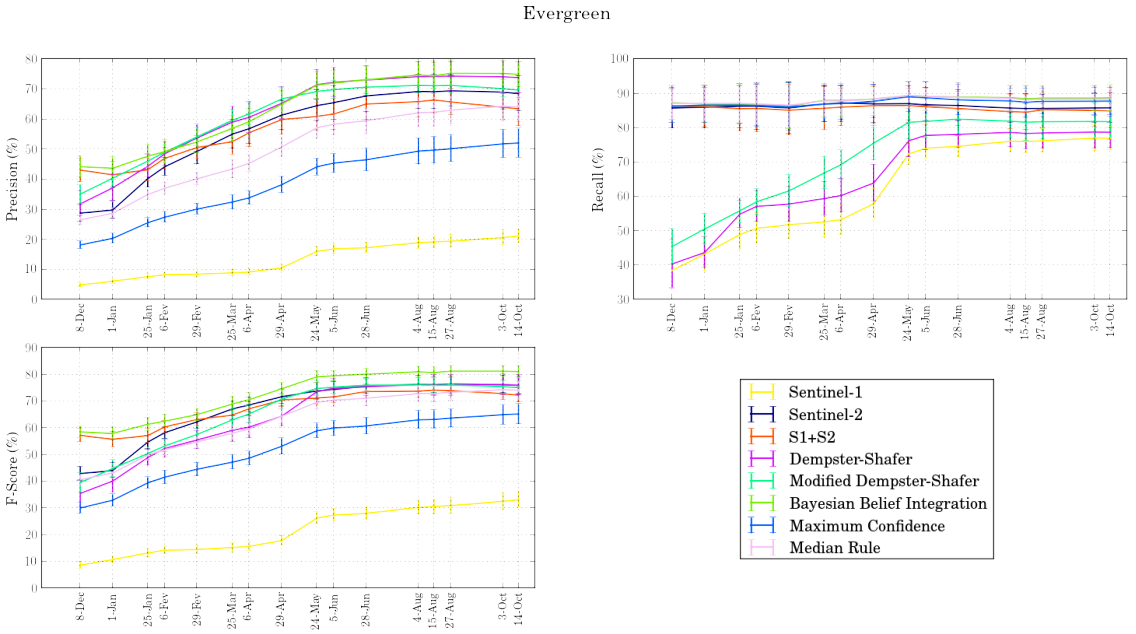


Figure 7.5: Temporal evaluation metrics. Precision, Recall and F-Score results along the agricultural season for the *Evergreen* class.

Regarding the proposed fusion strategies, BB method show the highest values for the three metrics. It is followed by the DS method which it is only outperformed by BB fusion for the precision values but it obtains one of the worst recall results. Nonetheless, it shows the second highest F-Score values. It should be remarked how the BB fusion is outperforming the other classifications without being affected by the S1 classification performance. This implies that the BB method is a robust fusion stage.

It is important to comment how S1S2 classification is performing. The temporal metrics show how the results increase and finally decrease. As mentioned before, this could be given due to the *Hughes phenomenon* explained in Section 3.1.

Finally, the performance of the MC strategy is commented. As it can be seen, MC fusion obtains the second worst F-Score results. Also, it shows low precision values along the season. This may be because the S1 model is predicting a high number of FP samples. Also, these samples obtain significant probability values. Therefore, the MC fusion performance is being affected by the significant confusions of the S1 classifier (See Appendix B for more details).

Sorghum class metrics

Finally, the last class studied is presented. In Figure 7.6 is detailed the temporal evaluation of the three proposed metrics for the *Sorghum* class. It can be seen that *Sorghum* class obtains one of the worst metric results. This fact it can be explained by the strong similarity existing between maize and sorghum crop classes. In fact, the S1 and S2 models present important confusions between all the summer crops.

Besides, by looking at the probability histogram of this class (See Figure A.10), it shows a large number of FP samples with high probability values. This large amount of FP samples explains the low precision values for S1 and S2 classifiers.

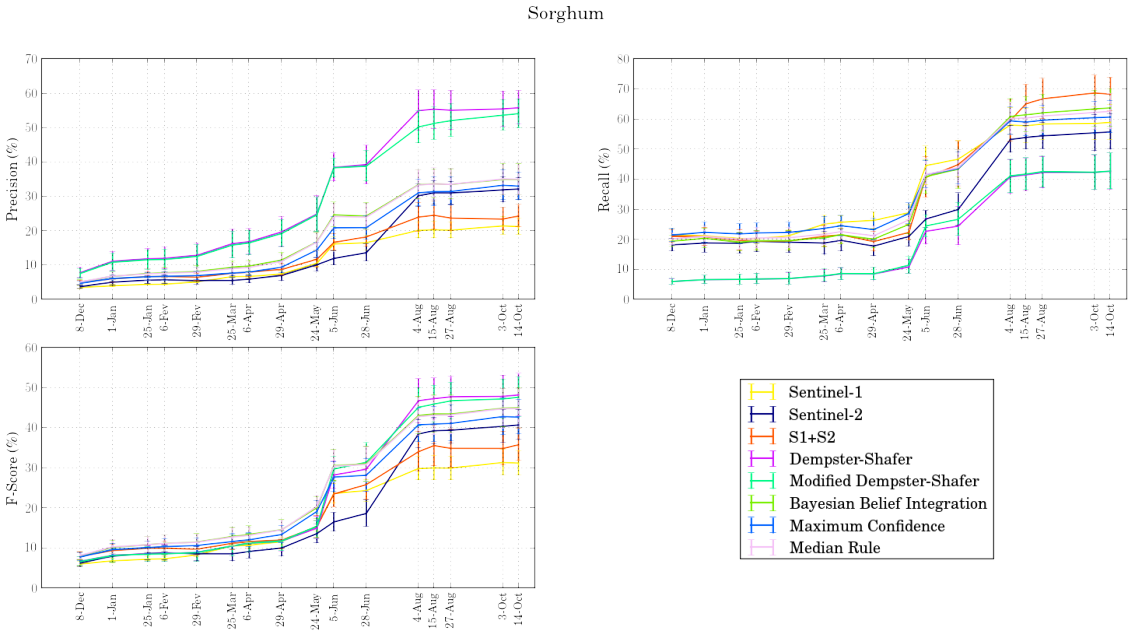


Figure 7.6: Temporal evaluation metrics. Precision, Recall and F-Score results along the agricultural season for the *Sorghum* class.

In this work, the interest of studying the *Sorghum* class is given by 1) the low classification performance and 2) the good results shown by DS methods.

Regarding the results obtained by the fusion strategies it can be seen that DS methods are able to obtain the highest precision and F-Score values. But, as shown in former cases, these methods obtains the lowest recall values. Therefore, the good performances presented by the DS methods are explained by the good precision results existing a significant difference between them and the other strategies.

As commented before, the S1 and S2 models are predicting summer crop samples as sorghum and these misclassified samples obtains high probability values. For this reason, the S1 and S2 classifications obtains poor precision results and also, the probabilistic methods are not able to obtain good results.

Finally, it should be remarked the strategies results over time. As explained before, the sorghum is a summer crop therefore the classification accuracies begin to improve as summer starts.

7.2 Evaluating the class confusion improvement

This study is carried out by comparing the S1 and S2 classification predictions with the fusion results. This comparison is performed using the confusion matrix results.

By means of the confusion matrix it may be possible to obtain a better idea of what the classification model is getting right and what types of errors it is making. Then, the goal of this study is to evaluate which class confusions might decrease by the use of a fusion stage at decision level.

Besides, an evaluation aiming at computing the agreements between classification strategies is presented. Also, a visual evaluation of these agreements is shown. The goal of this evaluation is to detect those pixels that are wrong predicted by the S1 and S2 classification but well predicted by the fusion strategy.

For simplicity, this study only presents the confusion analysis for the BB fusion technique. This fusion strategy is studied here because it has obtained the best classification results in the previous evaluation.

7.2.1 Analysis of the confusions between classes

A visual evaluation of the misclassified pixels is presented here. The study of the confusions between classes is shown for the *Shrubland* and *Orchard* classes since they present significant improvements for the BB fusion. Besides, this study is also carried out for different instants of time. But there are two different ways in which this analysis can be approached.

On the hand, confusions from FP samples are studied here. As mentioned in Section 3.4, the set of predicted pixels for a given class can be divided into TP and FP samples. Therefore, the interest of this study is to identify the proportion of falsely predicted pixels and to show to what classes belong.

On the other hand, confusions from FN samples are studied too. The set of reference pixels belonging to a given class can be divided into TP and FN samples. Therefore, the interest of this study is also to identify the proportion of reference pixels that are incorrectly predicted and to show to what classes belong.

Identifying the proportion and the classes of the misclassified pixels along the season allows to study 1) the similarities between classes, 2) how the classification models are learning and 3) how the fusion methods improve the S1 ans S2 classifications results.

Figure 7.7 and Figure 7.8 present the study for the confusions between classes for the *Shrubland* class. As explained before, there are two ways of handling this study. Then, upper figure corresponds to the FP approach while bottom figure corresponds to the FN approach.

As it may be seen, the figures are composed by three different plots. Each plot corresponds to the class confusions for the S1, S2 and BB strategies. For instance, if we look at plot on the left from Figure 7.7 it can be seen a set of bars for different instants of time. Each bar represents the total share (in %) of predicted pixels by S1 classifier for the *Shrubland* class. As explained before, this set of pixels can be divided into TP and FP pixels. Therefore, we are able to observe the % of pixels

well and falsely predicted. Also, it is possible to identify to which classes belongs the misclassified pixels.

In Figure 7.7 it can be seen that exists a trend to predict *Grassland*, *Deciduous* and *Fallow* samples as *Shrubland*. All of them are permanent classes. Therefore, comparing the plots corresponding to each strategy, it can be seen how the BB fusion is able to reduce FP confusions and to present the largest proportion of TP samples.

Comparing results shown in Figure 7.7 and Figure 7.8 it may be seen that *Shrubland* class classification implies a higher number of FP confusions. This can be noticed in upper figures since the proportion of TP samples (*i.e.* dark blue) shows a minor presence.

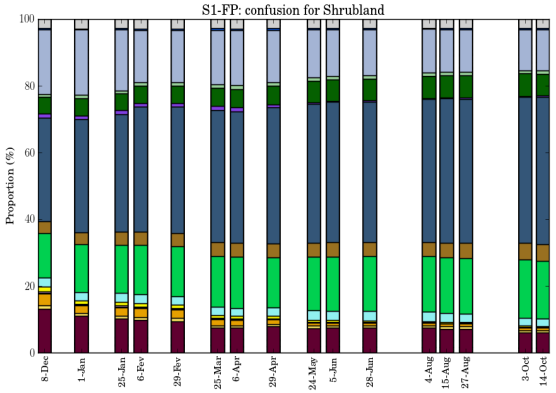
Following the study for the *Orchard* class is presented. Figure 7.9 and Figure 7.10 show the confusion plots for the S1, S2 and BB strategies for FP and FN approaches, respectively.

By looking at Figure 7.9, it is interesting to remark how the three classifications present a trend to predict *Grassland*, *Fallow* and *Vine* samples as *Orchard*. Comparing the three plots, it can be seen how the S1 classifier shows a large confusion for *Grassland* and *Vine* classes presenting the worst performance. The S2 and BB strategies present similar results where they also show a large proportion of *Grassland* samples. Nonetheless, BB fusion is able to reduce FP confusions for *Maize*, *Straw* or *Alfalfa* classes.

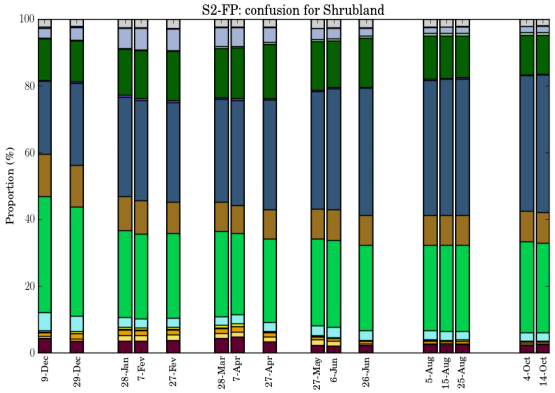
Another interesting remark it may be seen comparing the S1 and BB performances in Figure 7.9. The BB strategy is able to reduce a large proportion of *Vine* samples but, in contrast, it slightly increase the proportion of *Grassland* samples. This fact can be explained by means of the Grassland probability histogram (See Figure A.6). This figure presents FN samples with high probability values. Therefore, this fact can imply a limitation for the BB fusion.

Comparing results shown in Figure 7.9 and Figure 7.10 it may be seen that *Orchard* class classification implies a higher number of FP confusions. Therefore, this implies that *Orchard* class will obtain higher recall values. Besides, it is important to remark the different performances shown by S1 and S2 classifications where S2 strategy present almost twice the proportion of TP samples for both figures.

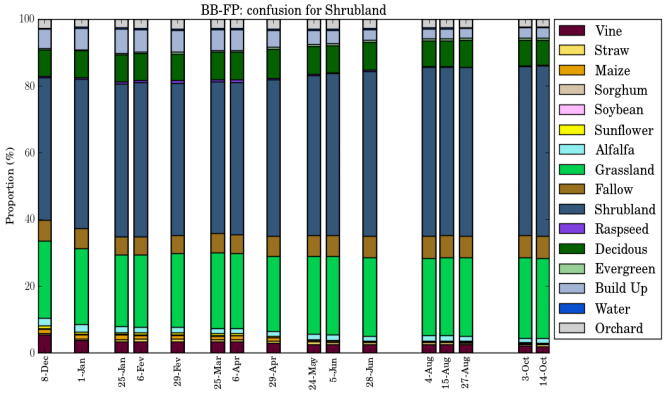
It may be seen that the BB fusion is able to reduce the proportion of misclassified samples despite the performance of the S1 classification. Nonetheless, this strategy does not show significant improvements.



(a) Sentinel-1

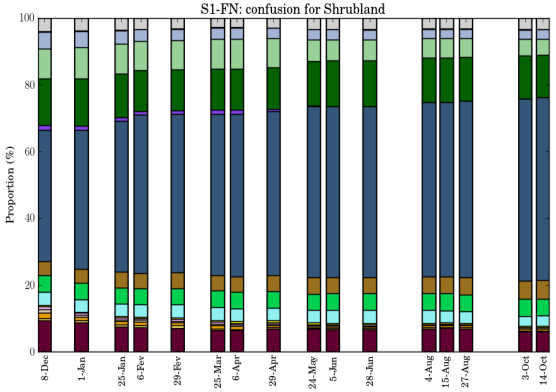


(b) Sentinel-2

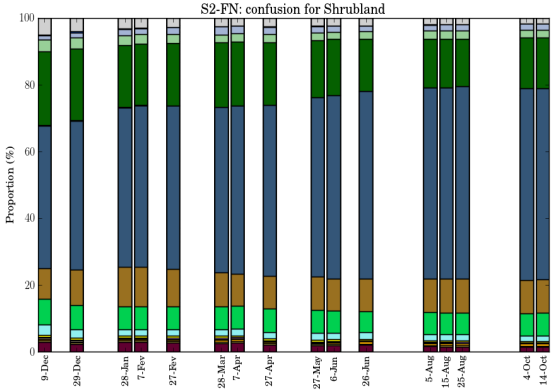


(c) Bayesian Belief Integration

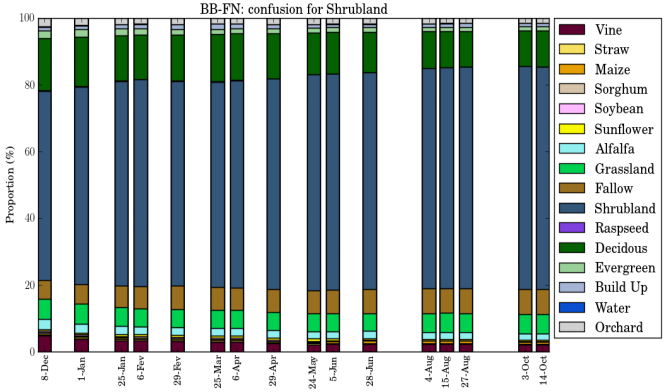
Figure 7.7: FP confusion approach for the *Shrubland* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.



(a) Sentinel-1

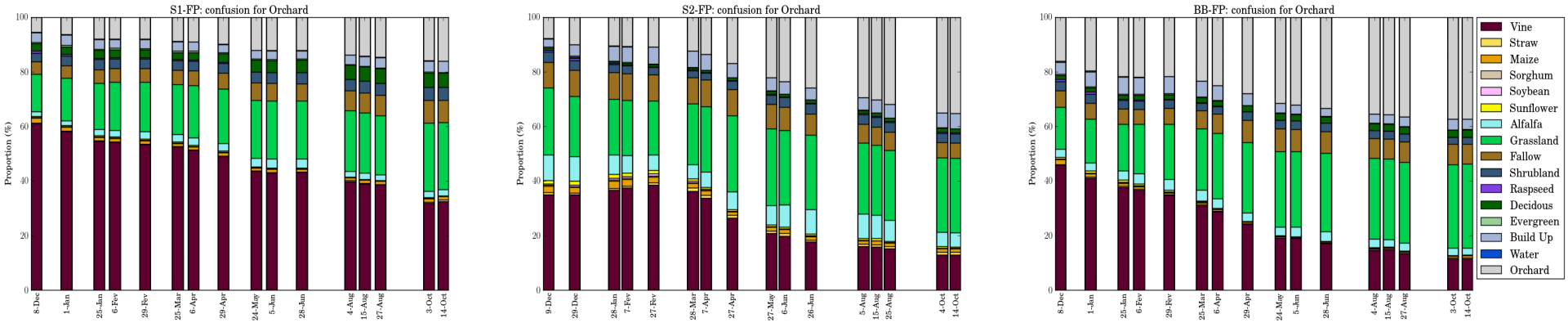


(b) Sentinel-2



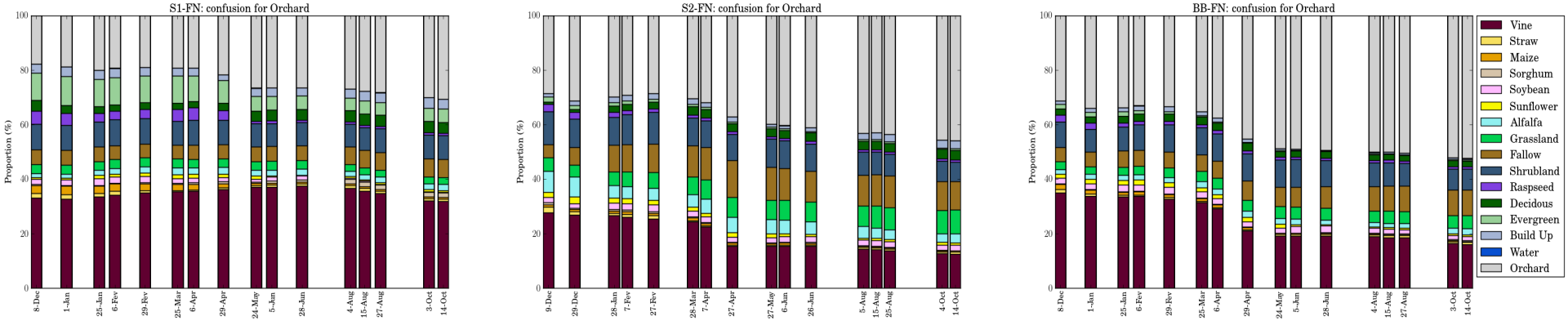
(c) Bayesian Belief Integration

Figure 7.8: FN confusion approach for the *Shrubland* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.



(a) Sentinel-1 (b) Sentinel-2 (c) Bayesian Belief Integration

Figure 7.9: FP confusion approach for the *Orchard* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.



(a) Sentinel-1 (b) Sentinel-2 (c) Bayesian Belief Integration

Figure 7.10: FN confusion approach for the *Orchard* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.

7.2.2 Analysis of the classifications agreements

A further way to study the performances of the proposed fusion techniques is studying the prediction agreements that the different classification strategies can have. Then, to perform this study, two different results are analyzed here. First of all, a statistical evaluation of the prediction agreements by classes is presented. As follows, a visual evaluation is performed by plotting the agreement information on a map.

As mentioned before, a visual evaluation does not provide information about the classifiers performance. Nonetheless, it may help to detect the pixel areas where the different strategies presents more disagreements. Besides, it may be useful to detect errors in the reference data.

In this work, the study of different agreement cases is considered. The study shall be able to show, clearly, the agreements and disagreement between the different strategies. Then, the interest of this study, is to identify those areas where the fusion methods can correctly perform despite the S1 or S2 strategies cannot. Thus, the different cases are categorized in the following way:

(i) $S1_{ok}, S2_{ok} \mid F_{ok}$: Proportion of pixels correctly classified for S1, S2, and the Fusion.

(ii) $S1_{ok}, S2_{ok} \mid F_{ko}$: Proportion of pixels correctly classified for S1 and S2 when the Fusion is wrong.

(iii) $S1_{ok}, S2_{ko} \mid F_{ok}$: Proportion of pixels correctly classified for S1 and the Fusion when S2 is wrong.

(iv) $S1_{ok}, S2_{ko} \mid F_{ko}$: Proportion of pixels correctly classified for S1 when S2 and the Fusion are wrong.

(v) $S1_{ko}, S2_{ok} \mid F_{ok}$: Proportion of pixels correctly classified for S2 and the Fusion when S1 is wrong.

(vi) $S1_{ko}, S2_{ok} \mid F_{ko}$: Proportion of pixels correctly classified for S2 when S1 and the Fusion are wrong.

(vii) $S1_{ko} = S2_{ko} \mid F_{ok}$: Proportion of pixels correctly classified for the Fusion when S1 and S2 are wrong as well, but with same label.

(viii) $S1_{ko} = S2_{ko} \mid F_{ko}$: Proportion of pixels incorrectly classified for the Fusion when S1 and S2 are wrong as well, but with same label.

(ix) $S1_{ko} \neq S2_{ko} \mid F_{ok}$: Proportion of pixels correctly classified for the Fusion when S1 and S2 are wrong as well, but with different label.

(x) $S1_{ko} \neq S2_{ko} \mid F_{ko}$: Proportion of pixels incorrectly classified for the Fusion when S1 and S2 are wrong as well, but with different label.

The study has been performed for all the fusion methods. However, for the sake of simplicity, only results involving the BB fusion strategy is presented. All the studied cases might be found in Appendix B.

A statistical evaluation of the classifications agreements

Here, a evaluation by classes is given for the resulting agreements. Table 7.4 presents the obtained results. The first important conclusion can be observed by looking at the first column. If the S1 and S2 classifiers succeed, then is highly probable that the fusion stage will correctly predict. Accordingly, if both single classifications are wrong, the fusion step is not a solution.

It is important to remark that the agreement between S2 classifier and BB fusion is higher than the agreement between S1 and BB strategies. However, the *Sorghum*, *Soybean* and *Deciduous* classes are an exception. If we look at Table 7.2, they obtain a better performance for the S2 classification. But, they obtain higher recall values for the S1 strategy than the S2 case. Therefore, it can be explained due to the fact that this study evaluates how the classifications are predicting the reference samples which is another way to define the recall metric.

Another interesting observation it can be seen by looking the fourth and sixth columns. These columns show the proportion of pixels that are well classified by one single classifier ($S1_{ok}$ or $S2_{ok}$) and are incorrectly classified by the fusion stage. Here, the proportion of pixels is always lower that results shown by third and fifth columns where is given the fusion stage success case.

Regarding the eighth column, it can be seen the results when both single classifiers are wrong predicting the same class and the fusion stage also is wrong. This fact can be done because of both classifiers are wrongly assigning the highest probability to the same class. Therefore, the fusion stage is not able obtain a good result.

In contrast, in the ninth column results when single classifiers are wrong are also given. But, in this case, both classifiers predicts different classes and the fusion stage succeeds. This can be due to pixels with low margin values. Those pixels present probability vectors where the probabilities are quite distributed in different classes. Then, those pixels present a difficulty for a single classifier but the fusion stage is able to exploit it. Nonetheless, tenth column shows the case where single classifiers are wrong predicting different classes and the fusion stage also fails. In this case, the fusion stage is not able to exploit the probability vectors of the S1 and S2 classifications.

Finally, an interesting case to comment is the results shown for the *Orchard* class. Comparing these results in the first and fifth column, it can be seen how there is a higher proportion of pixels that are correctly predicted by only the S2 and BB strategies. This is a particular case since the other classes show the higher values for the results in first column.

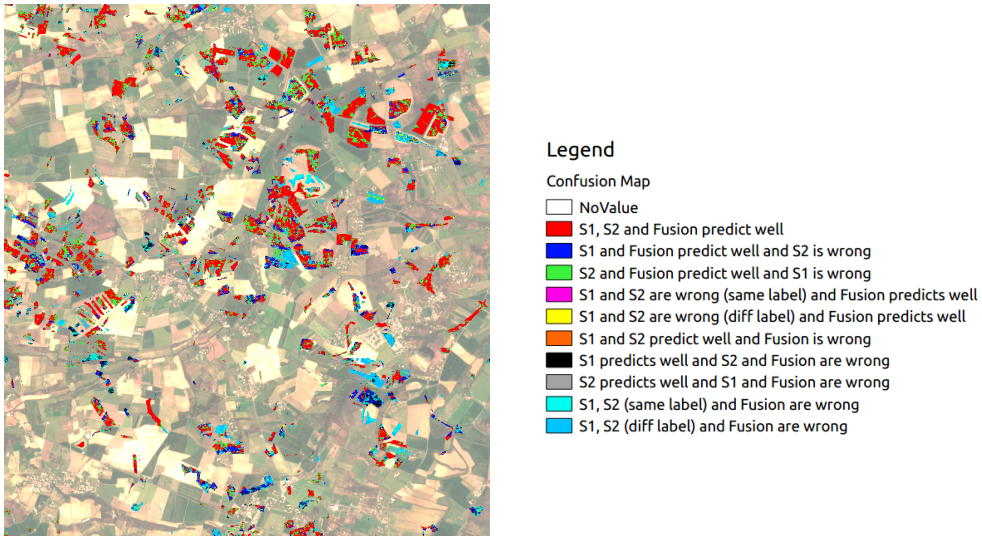
Class	$S1_{ok}S2_{ok}F_{ok}$	$S1_{ok}S2_{ok}F_{ko}$	$S1_{ok}S2_{ko}F_{ok}$	$S1_{ok}S2_{ko}F_{ko}$	$S1_{ko}S2_{ok}F_{ok}$	$S1_{ko}S2_{ok}F_{ko}$	$S1_{ko} = S2_{ko}F_{ok}$	$S1_{ko} = S2_{ko}F_{ko}$	$S1_{ko} \neq S2_{ko} \mid F_{ok}$	$S1_{ko} \neq S2_{ko} \mid F_{ko}$
Vine	83.28	0.00	4.72	1.44	6.03	0.60	0.00	1.46	0.51	1.97
Straw	86.81	0.00	3.14	0.50	4.48	0.81	0.00	2.44	0.21	1.62
Maize	72.26	0.00	5.58	2.29	8.73	1.00	0.00	5.36	0.89	3.90
Sorghum	42.71	0.00	11.19	5.08	7.48	5.66	0.00	10.25	2.43	15.20
Soybean	59.56	0.00	9.08	3.56	7.57	4.28	0.00	8.62	0.88	6.43
Sunflower	81.57	0.00	2.28	0.81	3.36	1.70	0.00	6.73	0.16	3.39
Alfalfa	44.08	0.00	11.24	6.44	11.41	4.15	0.00	10.29	2.40	10.00
Grassland	39.63	0.00	5.12	5.38	17.17	5.74	0.00	13.56	1.30	12.11
Fallow	35.19	0.00	8.96	7.35	19.45	4.43	0.00	11.15	1.83	11.63
Shrubland	38.60	0.00	11.29	4.31	14.97	4.88	0.00	10.89	2.24	12.83
Raspseed	73.74	0.00	3.49	0.30	1.03	0.83	0.00	14.70	0.06	5.86
Deciduous	78.27	0.00	6.09	1.50	5.07	2.10	0.00	2.90	0.35	3.71
Evergreen	69.60	0.00	4.03	3.40	14.71	1.55	0.00	2.79	0.40	3.52
Build up	76.47	0.00	3.25	0.60	14.18	1.87	0.00	0.73	0.69	2.21
Water	95.04	0.00	0.10	0.03	2.75	1.01	0.00	0.32	0.02	0.73
Orchard	19.08	0.00	5.81	4.74	23.57	4.67	0.00	18.33	5.09	18.71

Table 7.4: Statistical evaluation (in %) of the classification agreements for the S1, S2 and BB strategies.

A visual evaluation of the classifications agreements

Figure 7.11 presents a visual evaluation to study the spatial distribution of the classification agreements. At the beginnings of this section, ten different categories of agreements were presented. Then, this visual evaluation aims to plot in a map the explained agreements for the S1, S2 and BB classification results.

In Figure 7.11 three different maps are presented. These maps are given for three different dates concerning the beginnings, mid and the end of the agricultural season. As commented in Section 6.2, the maps are shown for different instants of time in order to study how the classification models are learning.



(a) 7-9th December

(b) 27-29th April

(c) 14th October

Figure 7.11: Agreement maps that show the classification agreements for the S1, S2 and BB strategies These maps are presented for three different dates along the season.

As it can be seen, these maps present the best agreement between the proposed

strategies at the end of the season (14th October). As commented in Section 6.2, the agreement maps become an useful tool to detect the areas where the models are mistaken. For instance, by looking on the bottom-left side of the studied area, it can be seen a particular area. This area presents that is being well predicted by the S1 and BB strategies at the beginnings of the season. But, as time goes by, the BB fusion starts to misclassified pixels. This can be explained due to the fact that the S2 classification present higher probability values. As explained in Section 6.2, the S2 classifier shows higher "confidence" level even if it is wrong.

7.3 Conclusions

This chapter has evaluated the proposed fusion strategies working at decision-level. The obtained results have shown the interest of performing the fusion stage by combining the output of RF models. The proposed fusion strategies have been able to outperform the classical single classification approach. Therefore, it is shown the interest of using more that one remote sensing data source. Also, the proposed fusion methods have been able to obtain better results than the classification approach consisting in use all the available remote sensing data as data input.

It has been shown the advantages of the fusion at decision level approach. The presented fusion strategies may divided into two categories, the fusion methods purely based on the use of the class probability vector and the fusion methods based on the Dempster-Shafer theory.

The DS category does not use the probabilistic RF output. In fact, this family only use the predicted label and the classifier accuracy information coming from the confusion matrix. The problem of this approach is that the same "fusion criterion" is applied for all the pixels predicted with a specific class. This can not be appropriate in some cases. Besides, the confidence of the classifier is not taking into account in the fusion task. For instance, imaging two pixels x and y predicted with the same class C_j. If the probability of the class C_j for the pixel x is 0.9 and the probability of the pixel y is 0.3, the fusion criterion does not should be the same. In contrast, a very interesting advantage of this approach is that it allow to weight the fusion decision according to the classification results of the single classifiers. For instance, if optical classifier obtains better results for one class, the predicted decision from the optical classifier will be important than the radar one.

Concerning the other fusion techniques, they use the RF classifier probabilities which allows us to exploit the uncertainty of the single classifiers. However, the fusion strategies studied here do not take into account the accuracies of the single classifiers. It implies than the optical and radar classifier decision have the same weight in the final prediction. It can suffer some limitations if the accuracy of the single classifiers is different. This problem can be seen in grassland and sorghum classes, looking at individual F-Score accuracies, the difference is 10%. The DS method take this consideration into account and it explains why is better than the other methods. In this case, the DS fusion method is giving more importance to the

decision of optical classifier prediction.

In order to study those fusion strategies, different evaluations have been presented in this chapter. These evaluations have shown how the BB fusion is able to improve the class confusions corresponding to the misclassified pixels. Results obtained from these evaluations increase the relevance of the class probability. The probabilistic fusion method may exploit the probabilities in order to improve the results. However, if the RF model misclassify a pixel with a high probability value the fusion stage is not able to improve the results.

It has been shown the importance of the class probability vector for the fusion stage. This study has presented how those fusion methods based on the handling of the class probabilities obtain the best results. Therefore, this chapter has corroborated the importance of a probabilistic fusion stage but also it has shown how this methods relies in the class probabilities values.

Chapter 8

New class probabilities estimation by using the Random Forest Out-Of-Bag error

As explained in Section 3.3, the classical class probabilities estimation done by RF is computed by using the ensemble of the tree classifiers. In fact, the probability that a sample has been assigned to a specific class corresponds to the number of trees in the forest that have vote for this class. Therefore, the class probability estimation is highly dependent on the accuracy of each learned tree and the diversity among them. But, counting the decisions of each tree as one vote it implies to assume that all the trees performs equally and have the same accuracy. Unfortunately, some trees in the forest can obtain worse accuracies given the complexity of data distribution in high dimensional space.

In practise, as shown in Equation 3.4, all trees have the same weight in class probability estimation. Therefore, the equal consideration of all the tree votes, can not be the most appropriate strategy to estimate the class probability of a sample. Note that a wrong estimation can be done if a large proportion of bad trees are included in the random forest.

Accordingly, the goal of the approach presented in this chapter is to reduce this negative effect on the class probability estimation. To perform it, a new estimation is proposed here taking into account the accuracy of each tree composing the random forest classier.

This chapter is structured as follows. Firstly, a section detailing the new estimation approach is given. Secondly, a section where an analysis of the OOB error for the radar and optical classifications is commented. Lastly, a experimental evaluation of the new approach is shown giving different metric analyzing the results.

8.1 Weighting the decision trees with the Out-Of-Bag error

As explained in Section 3.3, the Out-Of-Bag (OOB) error is a method that allows the measuring of the prediction error of the RF model. The prediction error may show how good is the classifier. In contrast, it totally depends on the training data and the model built.

Here, it is proposed a new approach of the probability vector that seeks to improve the RF model performance. The OOB error, estimated during the building of the model, can be seen as a metric of accuracy of each tree. Therefore, by means of the OOB error it is possible to assign different weights to each tree. Therefore, the idea of this approach consists in using the OOB error to modify the class probabilities estimation. Hence, those trees with higher accuracies will be more important in the decision process, and therefore, the probability values more accurate.

Hence, the new approach is given by redefining the Equation 3.4. In this context, consider a given sample x, where the kth-tree presents the OOB error ε_{tree_k}. The probability $p^k_{c_i}(x)$ that sample x belongs to class C_i for the kth-tree is redefined as:

$$p^k_{c_i}(x) = \begin{cases} (1 - \varepsilon_{tree_k}) & \text{If } argmax(m^{n_k(x)}) = C_i \\ 0 & \text{Otherwise} \end{cases} \tag{8.1}$$

where $m^{n_k(x)}$ contains the number of samples per class that fell on the leaf $n_k(x)$. Then, the new probability vector is estimated by counting this new parameter. In Figure 8.1, it can be seen the diagram of the new estimation approach.

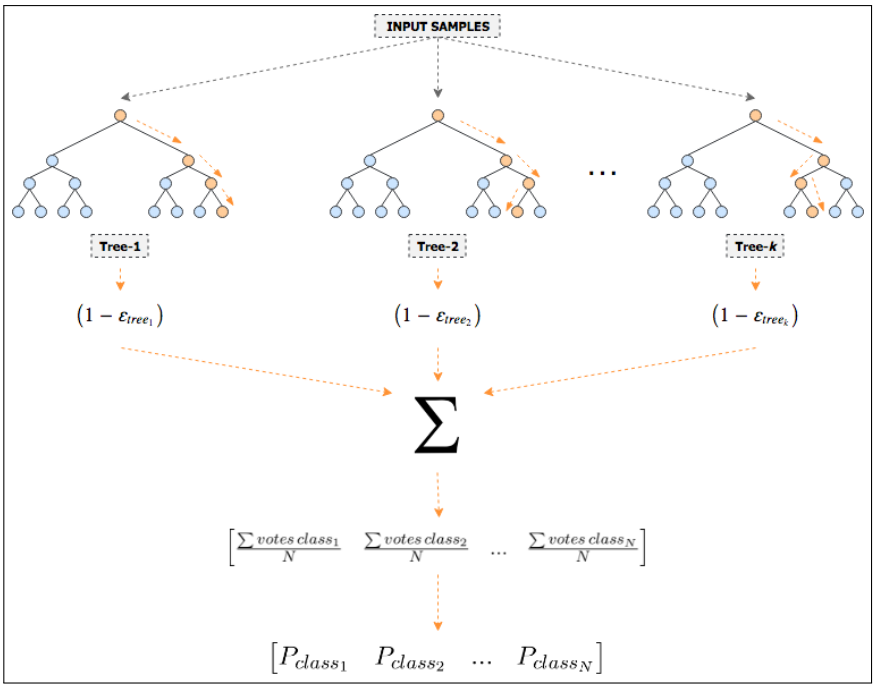


Figure 8.1: New class probability estimation by taking into account the accuracy of each individual tree.

8.2 Analysis of the Out-Of-Bag error

This section seeks to give a simple overview of the present OOB errors obtained during the training process of the single classifier. The training parameters for the learning proceeding can be found in Table 6.1. This parametrization has been done to carry out a fair comparison. Then, the OOB error will be obtained for $N = 100$ decision trees.

Figure 8.3 presents the ensemble of OOB errors of the optical and radar random forest classifiers. For each classifier, the individual error of the 100 trees composing the forest is shown. Besides, the studied learned models correspond to the classifiers trained with all the available dates. It means that they are the models obtained at 14h October. As it may be seen, the optical classifier obtain a low error value in comparison with radar model. Thus, it means that optical classifier is better than radar classifier which can be corroborated by looking the results of the previous chapters (see Chapters 6 and 7 for more details).

It is worthy mentioning the negative exponential behaviour that the OOB error presents meaning an unbalanced distribution of the OOB error along the trees. This particular performance of the OOB error deserves a deeper analysis. Unfortunately, this analysis has been carried out during the last internship weeks. Therefore, more efforts will need to be done in the future to understand this situation.

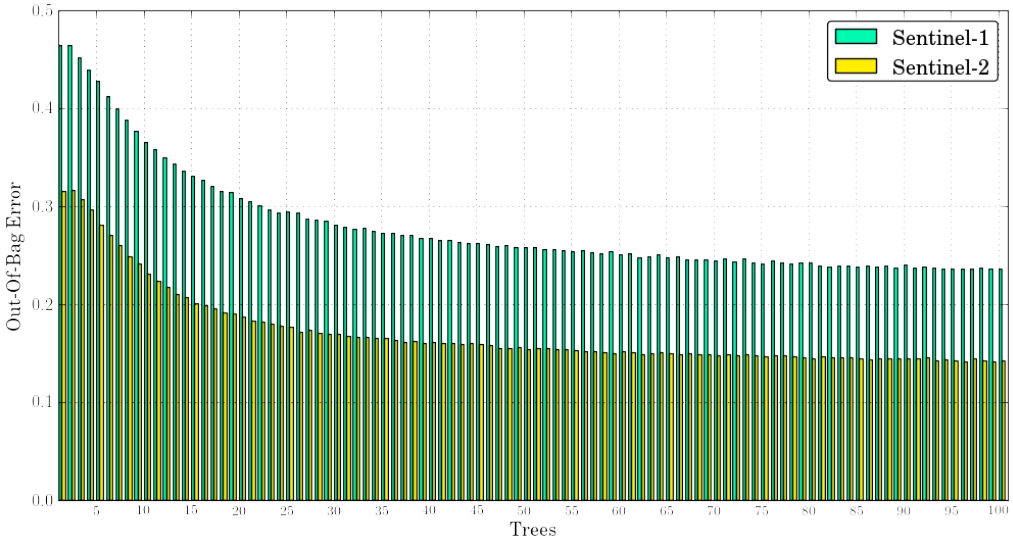


Figure 8.2: Out-Of-Bag error obtained during the training of the radar and optical models. The results are given for the 14th October.

In order to obtain a better understanding, Figure 8.3 shows the OOB error along the time for some specific trees. The selection of the studied trees has been randomly performed and it correspond to the threes number 1, 25, 50, 75 and 100.

As it can be observed, the OOB values decrease along the time for both classifier. This decrease is normal since more images are used to train the classifiers along the time. Therefore, the accuracy performances are improving along the year as shown in the results presented in the previous chapters. Besides, it can also observed in this figure that the OOB error of first tree is lower than the other trees along the time.

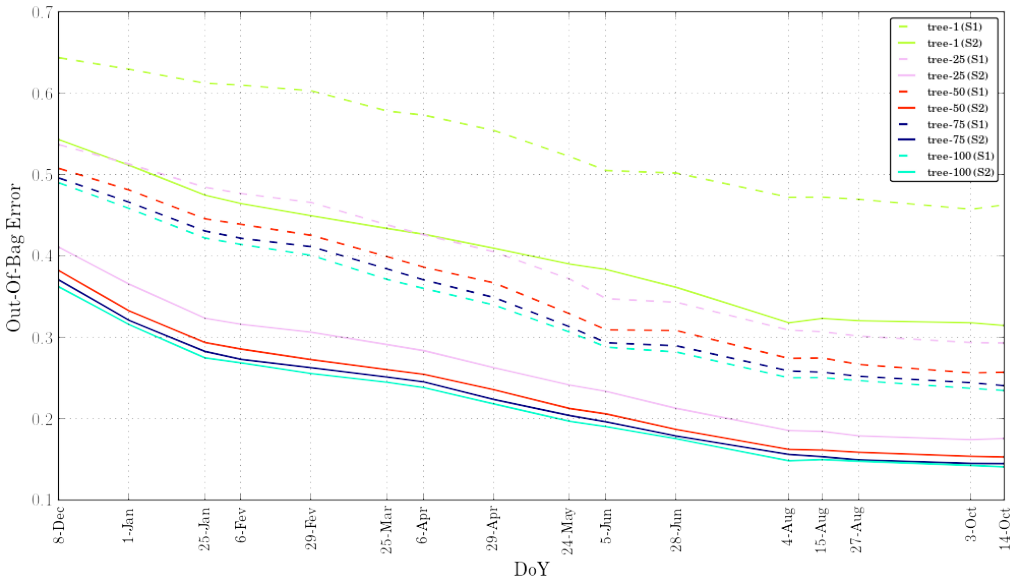


Figure 8.3: Temporal evolution of the OOB error during the training of the radar and optical models. The results are given for the trees 1, 25, 50, 75 and 100 of the ensemble.

8.3 Evaluation of the weighted class probability estimation

Once, the OOB error has been observed and analyzed, this section aims to show the performance results for the new probability vector estimation.

The classical approach of the probabilities implementation, takes advantage on the decision tree votes treating each tree as equiprobable. Therefore, this new implementation intends to improve the results by giving a particular weight to each tree directly related to its cross-validated accuracy.

In order to present the obtained outcomes this section is organized following the same division as Section 7.1. First of all, the Overall Accuracy is presented followed by the Precision, Recall and F-Score metrics. But unlike Section 7.1, and for simplicity, the classification strategies presented are not the same. For this case, the Dempster-Shafer-based methods has been neglected since they have shown irregular performance results. Following this reasoning, metrics are acquired for radar and optical single classifiers (S1 and S2) and the pure-probabilistic fusion techniques (BB, MR and MC).

8.3.1 Overall Accuracy

Figure 8.4 shows the corresponded plots of the OA for the proposed approaches. As it might be seen, this new estimation of the probabilities does not implies any improvement in general terms. Indeed, the involved changes are marginal where each presented approach achieves the same rate.

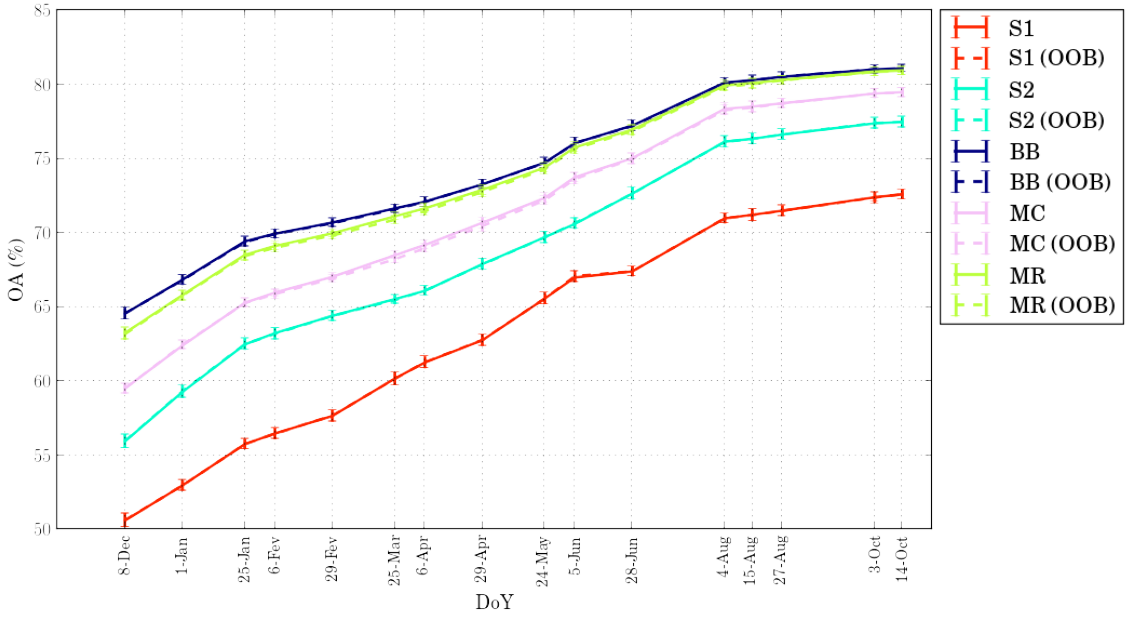


Figure 8.4: Influence of OOB error weighting in the OA metrics for the proposed classification strategies.

8.3.2 Precision, Recall and F-Score

Class metrics are presented in this section, in order to show a broader evaluation of the new development. Figure 8.4 has shown how the OOB error weighting does not involve any performance enhancement for this metric. But, minor variations could be noticed at some particular classes. For this reason, the following metrics are presented.

Evergreen metrics

Evergreen class are one of the classes that shows higher unbalance between S1 and S2 classifier behaviour. Nonetheless, it achieves significant results for the S2 strategy and the fusion methods.

In Figure 8.5 the results of the new approach for the proposed strategies are given. Besides, the former results are shown in order to obtain a better understanding. As might be appreciated in the precision plot, the MC and MR fusions present significant changes. This changes show how the results obtained by means of the new approach present higher values. Consequently, the F-Score metric also presents better results for these strategies. However, these fusions are the only ones that seems to show improvements. The *BB* fusion continues achieving best results for the former approach.

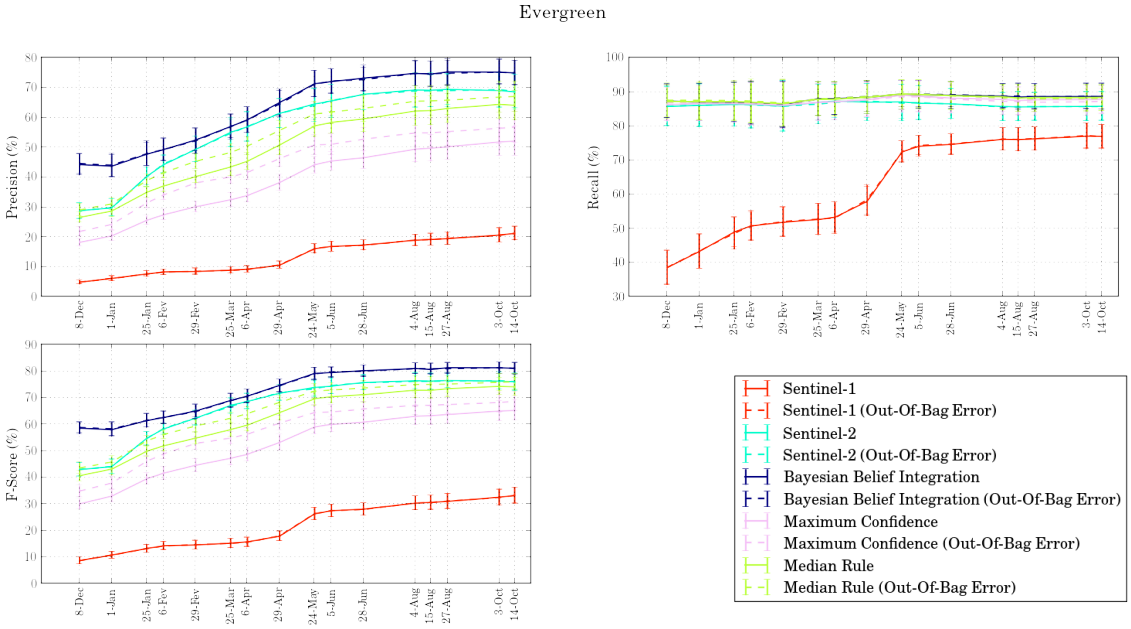


Figure 8.5: Precision, Recall and F-Score metrics for the *Evergreen* class. These results are shown for the presented probability estimation and the previous approach.

Shrubland metrics

Shrubland metrics are presented in Figure 8.6. This class is an interesting example since it present a behaviour completely contrary to the *Evergreen* case.

As it might be observed, *Shrubland* metrics plot similar results for S1 and S2 classifiers. In contrast, a great improvement for the pure-probabilistic fusion techniques can be appreciated. Regarding the results from the new approach, the only strategies that present different behaviour are the *MC* and *MR* fusion methods. But, oppositely to the *Evergreen* case, the new approach implies a worse performance.

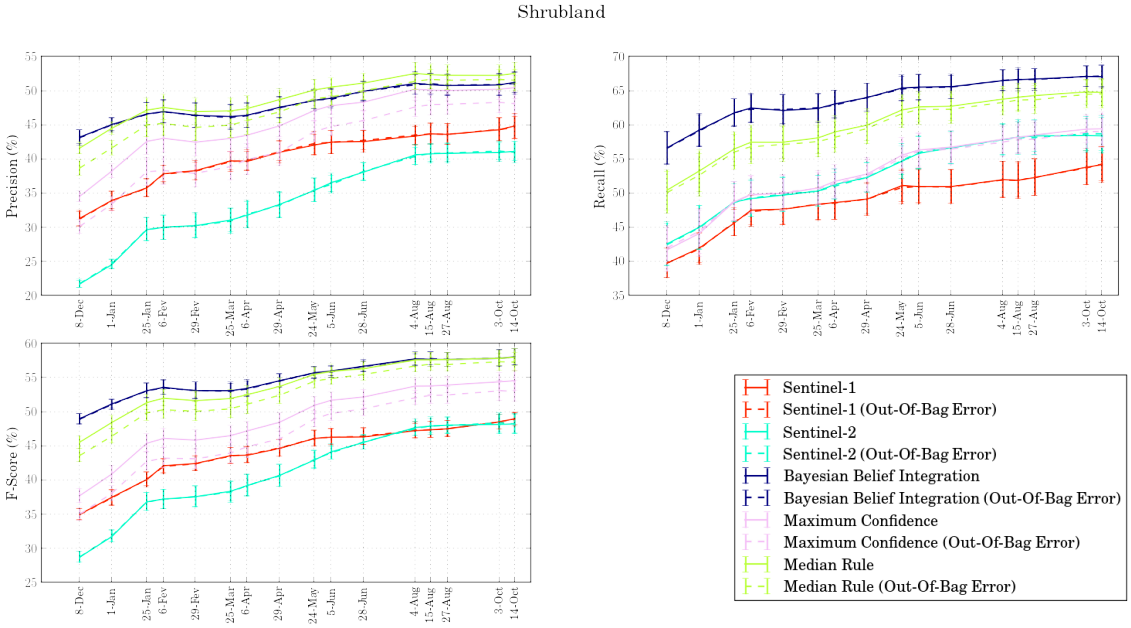


Figure 8.6: Precision, Recall and F-Score metrics for the *Shrubland* class. These results are shown for the presented probability estimation and the previous approach.

8.4 Conclusions

This chapter has presented a new strategy to estimate the class probability of unlabeled sample by using the RF structure.

Firstly, an analysis of OOB error has been shown. In this study, the estimated error has been given for the radar and optical single classifier. Besides, an temporal evolution of the error parameter has been described. The obtained results has shown how radar and optical classifiers obtain and estimated error related to their performance.

Secondly, the new class probabilities estimation has been tested for the S1, S2, BB, MC and MR strategies. The new classifications has been performed and evaluated. Besides, the results have been compared with the former metrics in Chapter 7. It has been show how the new approach may vary depending on the studied

class. Particularly, the MC and MR strategies have presented irregular behaviours corresponding to this new probabilities estimation. But, nonetheless, the overall results has been shown no significant improvement.

Therefore, it could be concluded that the proposed use of the OOB error metric as a weight, it does not has important advantages in the classification strategies. This fact it could be due to the limited variation of this parameter.

Part IV

Conclusions

Chapter 9

General Conclusions

9.1 Conclusions

This project has been carried out successfully, in the CESBIO facilities, in the framework of the SENSAGRI project.

The main goal of this work has been to define a classification strategy to achieve the joint use of the new optical and radar satellite images time series. The simplest solution to use both data sources consists into propose a fusion step at pixel level. This strategy consists in use all the available radar and optical data as input data. Unfortunately, the high dimension of the optical and radar data present some limitations in order to learned classification models. The important information can be lose in the high dimension and supervised classifier can have problems to detect discriminative information.

Accordingly, the proposed solution relies on a statistical fusion approach which allows us to combine the decision of single classifiers. The interest of the proposed fusion strategy is to improve classification accuracies by means of the probabilistic output from the RF model. The RF algorithm allow us to obtain the probabilities of belonging for a each class. Hence, this statistical output, instead of the classical label, enable us to get a better understanding of how the models are performing.

The classification strategies proposed in this work are developed for land cover mapping purposes. Therefore, first chapters have introduced some basic knowledge about land cover supervised classification and the input data used for the goals of the project.

The supervised classifier used in this study has been the Random Forest algorithm. This ensemble method has been proved to be suitable for the classification of time series by several studies [20, 37]. Criteria such as processing time, stability or robustness has been taken into account to select this classification algorithm. A description of this classifier has been presented in Chapter 3.

In the context of land cover mapping classification, the idea of combining the deci-

sions of several classifiers has been explored. The reasons for combining the outputs of multiple classifiers are compelling, because different classifiers may implicitly represent different useful aspects of a problem while no one classifier represents all useful aspects. One important consideration is that a class probability vector can be obtained as output of the RF model. Hence, the fusion step proposed here aims to exploit the use of these probabilities. These statistical fusion approaches have been compared with classical widely known combination methods such as the Dempster-Shafer theory. The ensemble of fusion approaches studied here have been presented in Chapter 5 where a theoretical description of the proposed fusion strategies has been detailed.

To better understand the probabilistic outputs of the RF classifier, a study has been carried out in Chapter 6. The goal of such study has been to study the relation between the class probability and the accuracy performance of the classifiers. It has allowed us to understand that some classes are predicted with higher "conviction" by the model than others. For this study, the different conclusions have been obtained by computing statistical and visual experiments.

The results obtained by the different fusion strategies have been presented and compared in Chapter 7. Several metrics has been proposed that have allow us to compare the proposed classification strategies. The results obtained in this chapter have shown that the addition of a fusion strategy at the decision level is the best approach in order to combine optical and radar information. This fusion stage has improved the simplest solution consisting in the pixel level fusion and the radar and optical single classifications. Several statistical and visual evaluations have been studied to highlight the good performances of fusion methods. These evaluations has shown how the fusion strategies are able to to decrease the confusions of the RF models. Besides, it has shown how fusion step may correctly predict those pixels misclassified by the single classifiers. Therefore, it has been proved that the fusion of classifiers decision is able to achieve better results than other strategies. Lastly, the fusion strategy that has obtained the best results is the true probabilistic approach called Bayesian Belief Integration.

Chapter 8 has presented a new class probability estimation for the Random Forest classifier. This approach is based on the use of a weighting step taking into account the accuracy of the individual trees in the forest. The interest of this strategy was to give more relevancy to the trees achieving the best accuracies. In this approach, the weight of each tree is proportional to the Out-Of-Bag error, a prediction error calculated for each tree during the training of the model. The use of the new weighted class probability vectors have been tested on the fusion strategies presented in Chapter 8. Several metric has been presented for different classes. Besides, a study of the OOB error of the single classifiers has been given. The obtained results have not achieve a significant improvement. However, it should be remarked that these experiments were carried out the last weeks of this internship. Therefore, it ca be an interesting topic to research in the future.

This research is the important improvement achieved by statistical fusion techniques. Such strategies have demonstrated their interest in a real application in order to exploit jointly radar and optical satellite imagery. Therefore, the good re-

sults have motivated to include the fusion strategy in to the operational processing chain prototype of the H2020 project.

9.2 Further development

The perspectives of this work are multiples. In order to obtain a better accuracy in future classification strategies, the future works shall involves a greater knowledge of the probabilities coming from each decision tree of the forest. In other words, a more precise probability vector may be obtained. Hence, a deeper digging in the RF probabilities estimation is needed.

It has been proved that fusion methods based on the RF probabilities obtain better accuracies. In that way, the performance of the ensemble of trees might be improved. A new weighting approach can be studied in order to determine those trees with highest accuracies. Hence, as mentioned before, a new estimation of the class probability vector can be developed. [20] proposed how to exploit the RF algorithm in order to obtain a new class probability vector. Generally, every tree of the ensemble votes 0 or 1 on the decision process. But, instead of a vote, each tree can output a class probability vector. Therefore, a new weighting step may be implemented using those probabilities.

the individual optical and radar classification performances can be improved. Some experiments have shown that the addition of spectral indices or ratio bands can improve the RF accuracies. For instance,...

Moreover, the individual optical and radar classification performances can be improved. Some experiments have shown that the addition of spectral indices or ratio bands can improve the RF accuracies. For instance, the Normalized Difference Vegetation Index (NDVI) or the ratio primitive $\frac{VH}{VV}$ can be added for the optical and radar classifications, respectively.

As it can be seen, the performance improvement of a classification framework could be cope from different angles and only trying to exploit the best from each block of the processing chain will be possible to meet the best accuracies.

Bibliography

[1] Johannes J Feddema, Keith W Oleson, Gordon B Bonan, Linda O Mearns, Lawrence E Buja, Gerald A Meehl, and Warren M Washington. The importance of land-cover change in simulating future climates. *Science*, 310(5754):1674–1678, 2005.

[2] Billie L Turner, Eric F Lambin, and Anette Reenberg. The emergence of land change science for global environmental change and sustainability. *Proceedings of the National Academy of Sciences*, 104(52):20666–20671, 2007.

[3] Andreas Wacke. Fritz sturm 1929-2015. *Zeitschrift der Savigny-Stiftung für Rechtsgeschichte, Romanistische Abteillung*, 133:648–653, 2016.

[4] David M Olson, Eric Dinerstein, Eric D Wikramanayake, Neil D Burgess, George VN Powell, Emma C Underwood, Jennifer A D'amico, Illanga Itoua, Holly E Strand, John C Morrison, et al. Terrestrial ecoregions of the world: A new map of life on earth: A new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *BioScience*, 51(11):933–938, 2001.

[5] Shannon M Sterling, Agnès Ducharne, and Jan Polcher. The impact of global land-cover change on the terrestrial water cycle. *Nature Climate Change*, 3(4):385, 2013.

[6] Tobias Kuemmerle, Karlheinz Erb, Patrick Meyfroidt, Daniel Müller, Peter H Verburg, Stephan Estel, Helmut Haberl, Patrick Hostert, Martin R Jepsen, Thomas Kastner, et al. Challenges and opportunities in mapping land use intensity globally. *Current opinion in environmental sustainability*, 5(5):484–493, 2013.

[7] Stacy L Ozesmi and Marvin E Bauer. Satellite remote sensing of wetlands. *Wetlands ecology and management*, 10(5):381–402, 2002.

[8] Cristina Gómez, Joanne C White, and Michael A Wulder. Optical remotely sensed time series data for land cover classification: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 116:55–72, 2016.

[9] Huiran Jin, Giorgos Mountrakis, and Stephen V Stehman. Assessing integration of intensity, polarimetric scattering, interferometric coherence and spatial texture metrics in palsar-derived land cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 98:70–84, 2014.

[10] Wayne S Walker, Claudia M Stickler, Josef M Kellndorfer, Katie M Kirsch, and Daniel C Nepstad. Large-area classification and mapping of forest and land cover in the brazilian amazon: A comparative analysis of alos/palsar and landsat data sources. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 3(4):594–604, 2010.

[11] Hasi Bagan, Tsuguki Kinoshita, and Yoshiki Yamagata. Combination of avnir-2, palsar, and polarimetric parameters for land cover classification. *IEEE Transactions on Geoscience and Remote Sensing*, 50(4):1318–1328, 2012.

[12] S Erasmi and A Twele. Regional land cover mapping in the humid tropics using combined optical and sar satellite data—a case study from central sulawesi, indonesia. *International Journal of Remote Sensing*, 30(10):2465–2478, 2009.

[13] Yunlin Zhang, Kun Shi, Yongqiang Zhou, Xiaohan Liu, and Boqiang Qin. Monitoring the river plume induced by heavy rainfall events in large, shallow, lake taihu using modis 250 m imagery. *Remote sensing of environment*, 173:109–121, 2016.

[14] Zbyněk Malenovskỳ, Helmut Rott, Josef Cihlar, Michael E Schaepman, Glenda García-Santos, Richard Fernandes, and Michael Berger. Sentinels for science: Potential of sentinel-1,-2, and-3 missions for scientific observations of ocean, cryosphere, and land. *Remote Sensing of Environment*, 120:91–101, 2012.

[15] Ramon Torres, Paul Snoeij, Dirk Geudtner, David Bibby, Malcolm Davidson, Evert Attema, Pierre Potin, BjÖrn Rommen, Nicolas Floury, Mike Brown, et al. Gmes sentinel-1 mission. *Remote Sensing of Environment*, 120:9–24, 2012.

[16] M Drusch, U Del Bello, S Carlier, O Colin, V Fernandez, F Gascon, B Hoersch, C Isola, P Laberinti, P Martimort, et al. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote sensing of Environment*, 120:25–36, 2012.

[17] Jesús Delegido, Jochem Verrelst, Luis Alonso, and José Moreno. Evaluation of sentinel-2 red-edge bands for empirical estimation of green lai and chlorophyll content. *Sensors*, 11(7):7063–7081, 2011.

[18] Alyssa K Whitcraft, Inbal Becker-Reshef, and Christopher O Justice. A framework for defining spatially explicit earth observation requirements for a global agricultural monitoring initiative (geoglam). *Remote Sensing*, 7(2):1461–1481, 2015.

[19] Jordi Inglada, Marcela Arias, Benjamin Tardy, Olivier Hagolle, Silvia Valero, David Morin, Gérard Dedieu, Guadalupe Sepulcre, Sophie Bontemps, Pierre Defourny, et al. Assessment of an operational system for crop type map production using high temporal and spatial resolution satellite optical imagery. *Remote Sensing*, 7(9):12356–12379, 2015.

[20] Charlotte Pelletier. *Cartographie de l'occupation des sols à partir de séries temporelles d'images satellitaires à hautes résolutionsIdentification et traitement des données mal étiquetées*. PhD thesis, 12 2017.

[21] Neha Joshi, Matthias Baumann, Andrea Ehammer, Rasmus Fensholt, Kenneth Grogan, Patrick Hostert, Martin Rudbeck Jepsen, Tobias Kuemmerle, Patrick Meyfroidt, Edward TA Mitchard, et al. A review of the application of optical and radar remote sensing data fusion to land use mapping and monitoring. *Remote Sensing*, 8(1):70, 2016.

[22] David Small and Adrian Schubert. Guide to asar geocoding. *Issue*, 1(19.03):2008, 2008.

[23] J Bruniquel and A Lopes. Multi-variate optimal speckle reduction in sar imagery. *International journal of remote sensing*, 18(3):603–627, 1997.

[24] Olivier Hagolle, Sylvia Sylvander, Mireille Huc, Martin Claverie, Dominique Clesse, Cécile Dechoz, Vincent Lonjou, and Vincent Poulain. Spot-4 (take 5): simulation of sentinel-2 time series on 45 large sites. *Remote Sensing*, 7(9):12242–12264, 2015.

[25] Thomas R Loveland, Bradley C Reed, Jesslyn F Brown, Donald O Ohlen, Zhiliang Zhu, LWMJ Yang, and James W Merchant. Development of a global land cover characteristics database and igbp discover from 1 km avhrr data. *International Journal of Remote Sensing*, 21(6-7):1303–1330, 2000.

[26] Reza Khatami, Giorgos Mountrakis, and Stephen V Stehman. A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research. *Remote Sensing of Environment*, 177:89–100, 2016.

[27] Gordon Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE transactions on information theory*, 14(1):55–63, 1968.

[28] Giles M Foody and Ajay Mathur. The use of small training sets containing mixed pixels for accurate hard image classification: Training on mixed spectral responses for classification by a svm. *Remote Sensing of Environment*, 103(2):179–189, 2006.

[29] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[30] Pall Oskar Gislason, Jon Atli Benediktsson, and Johannes R Sveinsson. Random forests for land cover classification. *Pattern Recognition Letters*, 27(4):294–300, 2006.

[31] Jordi Inglada, Arthur Vincent, Marcela Arias, Benjamin Tardy, David Morin, and Isabel Rodes. Operational high resolution land cover map production at the country scale using satellite image time series. *Remote Sensing*, 9(1):95, 2017.

[32] Victor Francisco Rodriguez-Galiano, Bardan Ghimire, John Rogan, Mario Chica-Olmo, and Juan Pedro Rigol-Sanchez. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67:93–104, 2012.

[33] Mahesh Pal. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1):217–222, 2005.

[34] Pierre Lassalle, Jordi Inglada, Julien Michel, Manuel Grizonnet, and Julien Malik. A scalable tile-based framework for region-merging segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 53(10):5473–5485, 2015.

[35] Robin Genuer, Jean-Michel Poggi, Christine Tuleau-Malot, and Nathalie Villa-Vialaneix. Random forests for big data. *Big Data Research*, 9:28–46, 2017.

[36] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168. ACM, 2006.

[37] Charlotte Pelletier, Silvia Valero, Jordi Inglada, Nicolas Champion, and Gérard Dedieu. Assessing the robustness of random forests to map land cover with high resolution satellite image time series over large areas. *Remote Sensing of Environment*, 187:156–168, 2016.

[38] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

[39] Lior Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2):1–39, 2010.

[40] Gavin Brown, Jeremy Wyatt, Rachel Harris, and Xin Yao. Diversity creation methods: a survey and categorisation. *Information Fusion*, 6(1):5–20, 2005.

[41] Ludmila I Kuncheva and Christopher J Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 51(2):181–207, 2003.

[42] Andrew Mellor and Samia Boukir. Exploring diversity in ensemble classification: Applications in large area land cover mapping. *ISPRS Journal of Photogrammetry and Remote Sensing*, 129:151–161, 2017.

[43] Robert Bryll, Ricardo Gutierrez-Osuna, and Francis Quek. Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. *Pattern recognition*, 36(6):1291–1302, 2003.

[44] Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8):832–844, 1998.

[45] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[46] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[47] Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *Icml*, volume 96, pages 148–156. Bari, Italy, 1996.

[48] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[49] Robin Genuer, Jean-Michel Poggi, and Christine Tuleau-Malot. Variable selection using random forests. *Pattern Recognition Letters*, 31(14):2225–2236, 2010.

[50] Borut Sluban. *Ensemble-based Noise and Outlier Detection: Doctoral Dissertation*. PhD thesis, B. Sluban, 2014.

[51] Desheng Liu and Ruiliang Pu. Downscaling thermal infrared radiance for subpixel land surface temperature retrieval. *Sensors*, 8(4):2695–2706, 2008.

[52] Lorenzo Busetto, Michele Meroni, and Roberto Colombo. Combining medium and coarse spatial resolution satellite data to improve the estimation of subpixel ndvi time series. *Remote Sensing of Environment*, 112(1):118–131, 2008.

[53] DG Leckie et al. Synergism of synthetic aperture radar and visible/infrared data for forest type discrimination. *PE&RS, Photogrammetric Engineering & Remote Sensing*, 56(9):1237–1246, 1990.

[54] Jan Stefanski, Tobias Kuemmerle, Oleh Chaskovskyy, Patrick Griffiths, Vassiliy Havryluk, Jan Knorn, Nikolas Korol, Anika Sieber, and Björn Waske. Mapping land management regimes in western ukraine using optical and sar data. *Remote Sensing*, 6(6):5279–5305, 2014.

[55] Glenn Shafer. *A mathematical theory of evidence*, volume 42. Princeton university press, 1976.

[56] S. Le Hegarat-Mascle, I. Bloch, and D. Vidal-Madjar. Application of dempster-shafer evidence theory to unsupervised classification in multisource remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 35(4):1018–1031, Jul 1997.

[57] JW Guan and David A Bell. Generalization of the dempster-shafer theory. In *IJCAI*, pages 592–597, 1993.

[58] Jeffrey A Barnett. Calculating dempster-shafer plausibility. *IEEE transactions on pattern analysis and machine intelligence*, 13(6):599–602, 1991.

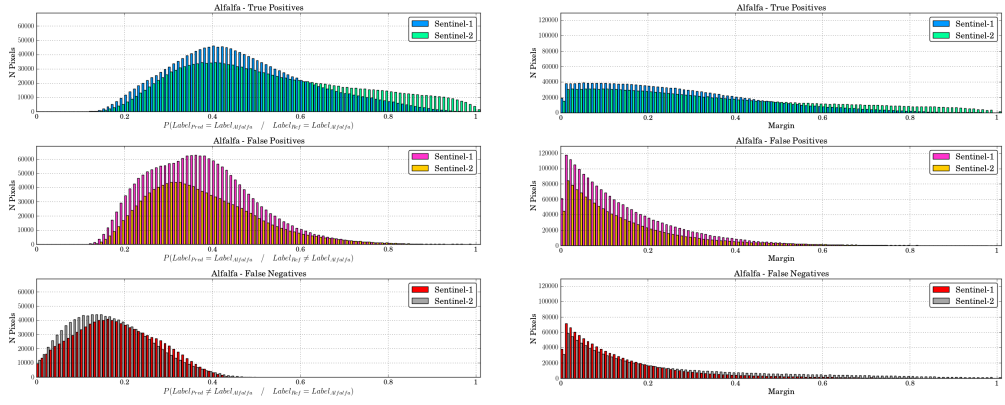
Appendices

Appendix A

Evaluation of the prediction results of the radar and optical single classifiers

The following appendix presents the results of the single classifiers predictions. These results have been statistically analyzed. Hence, the probability and margin histograms for the classified classes are given as follows.

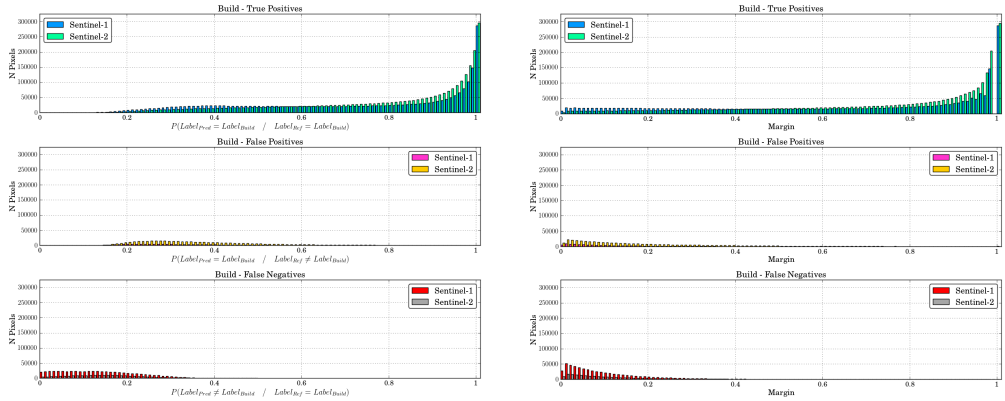
Alfalfa class results



(a) (b)

Figure A.1: Probability (a) and Margins (b) histograms for the Alfalfa class.

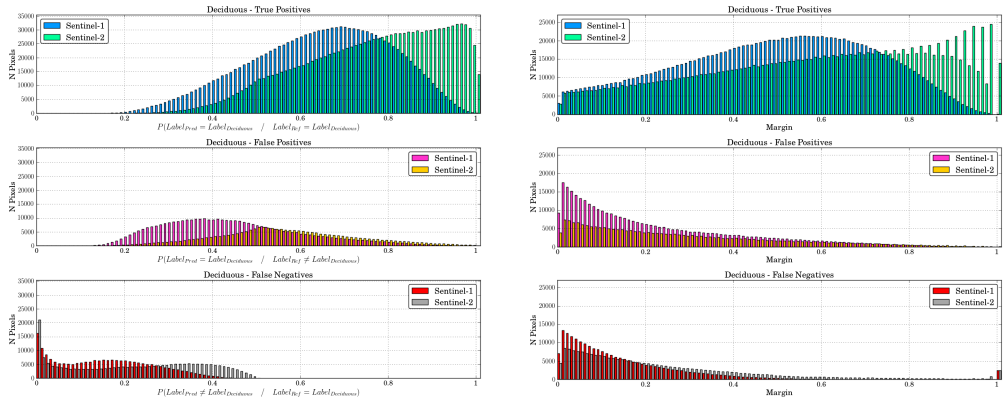
Build up class results



(a) (b)

Figure A.2: Probability (a) and Margins (b) histograms for the Build up class.

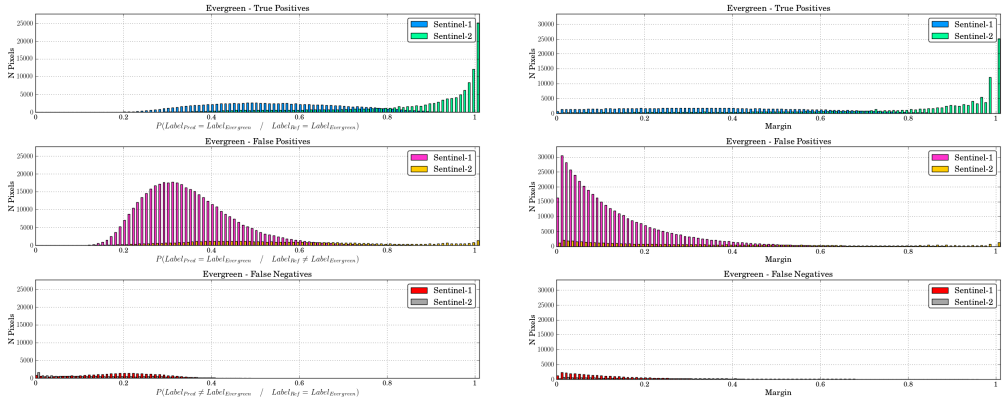
Deciduous class results



(a) (b)

Figure A.3: Probability (a) and Margins (b) histograms for the Deciduous class.

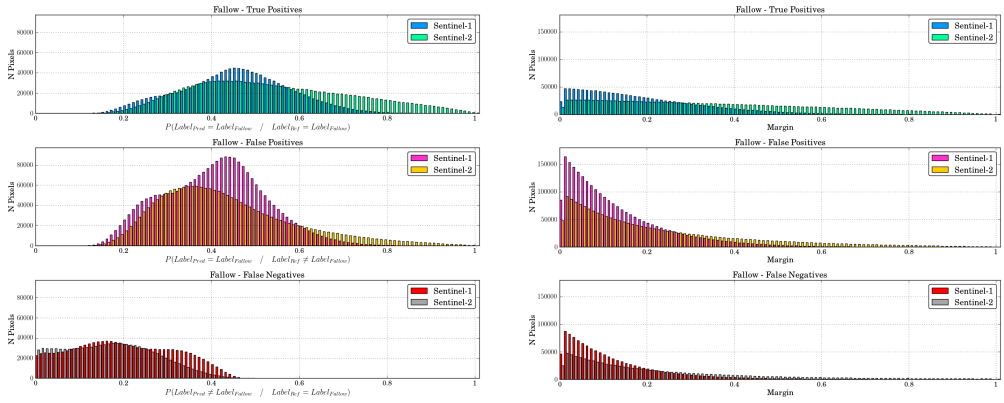
Evergreen class results



(a) (b)

Figure A.4: Probability (a) and Margins (b) histograms for the Evergreen class.

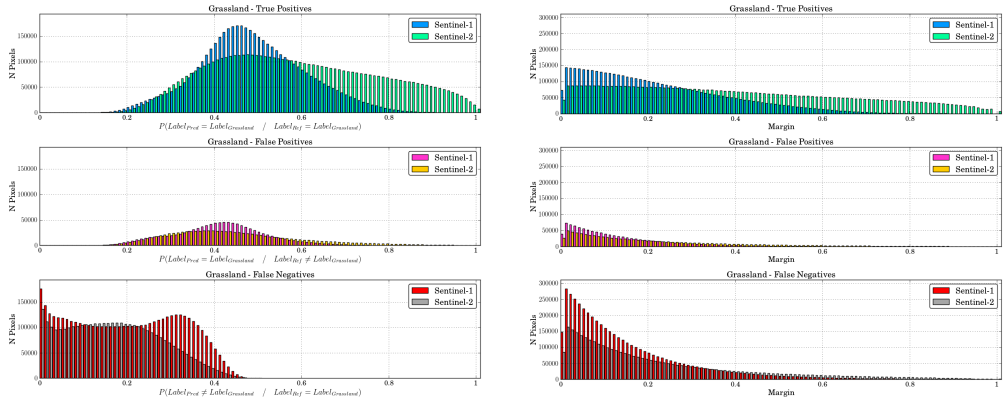
Fallow class results



(a) (b)

Figure A.5: Probability (a) and Margins (b) histograms for the Fallow class.

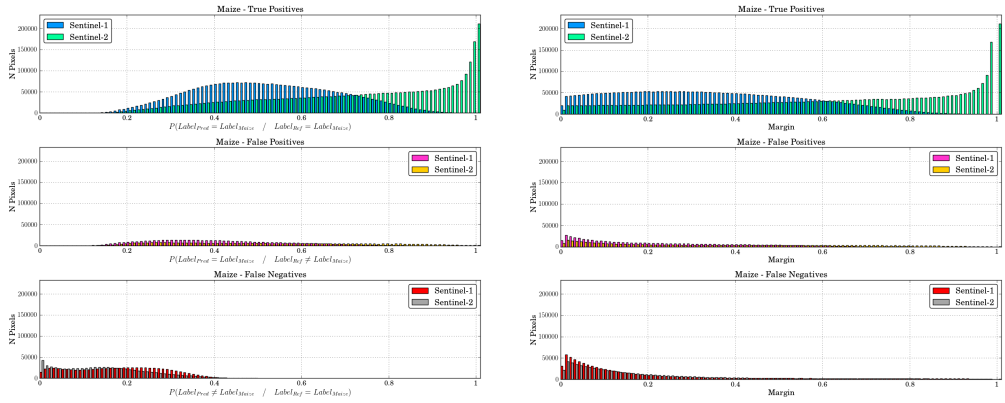
Grassland class results



(a) (b)

Figure A.6: Probability (a) and Margins (b) histograms for the Grassland class.

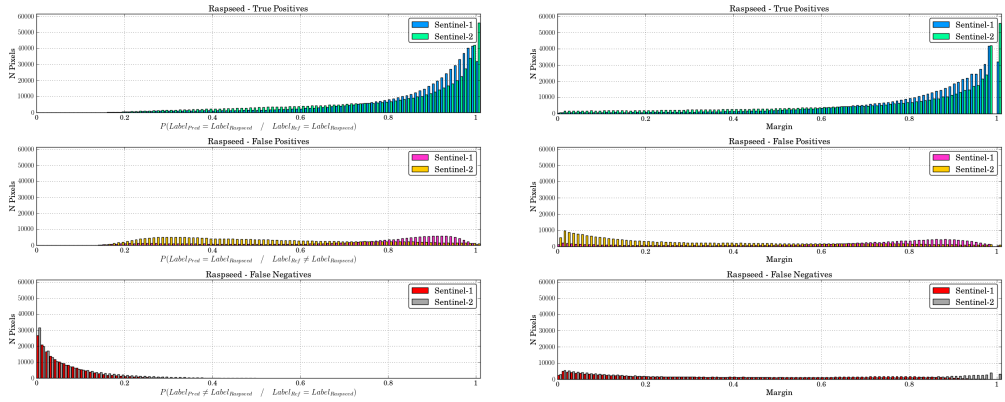
Maize class results



(a) (b)

Figure A.7: Probability (a) and Margins (b) histograms for the Maize class.

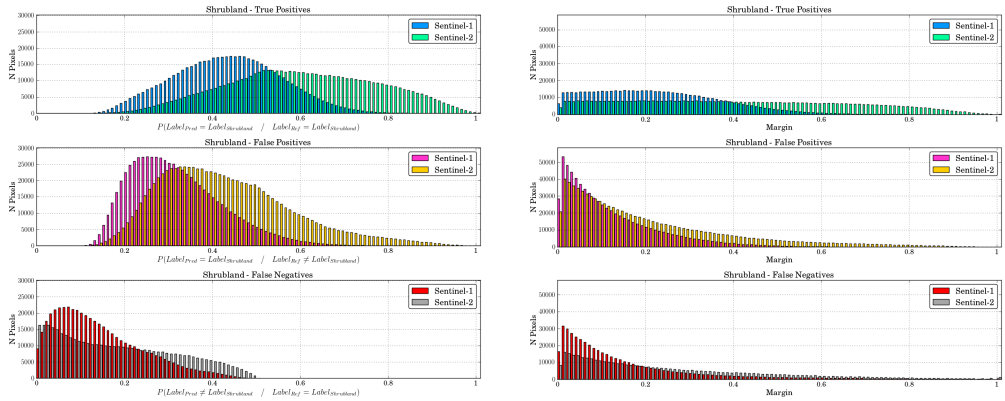
Rapeseed class results



(a) (b)

Figure A.8: Probability (a) and Margins (b) histograms for the Rapeseed class.

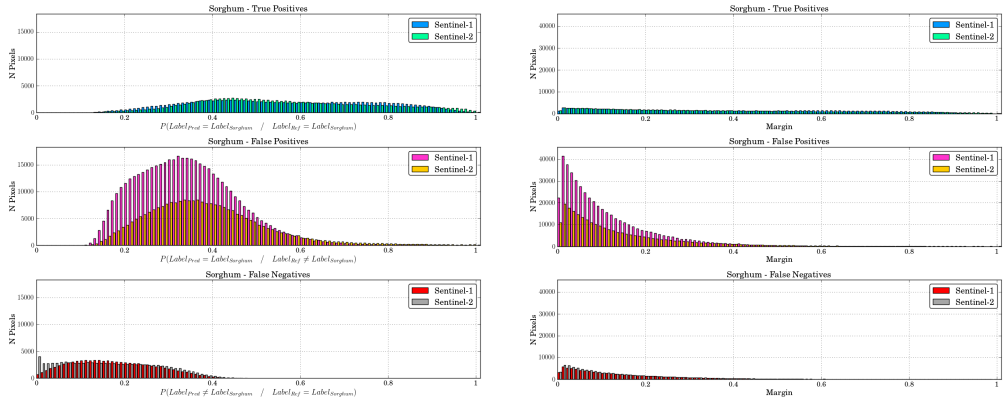
Shrubland class results



(a) (b)

Figure A.9: Probability (a) and Margins (b) histograms for the Shrubland class.

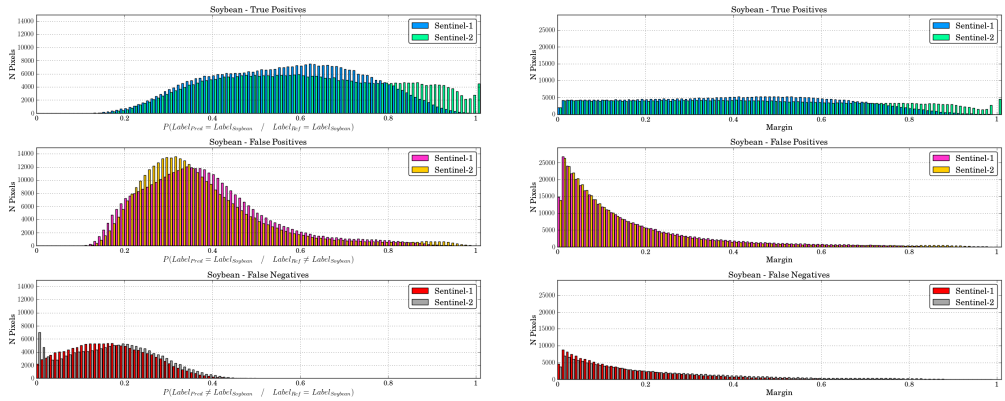
Sorghum class results



(a) (b)

Figure A.10: Probability (a) and Margins (b) histograms for the Sorghum class.

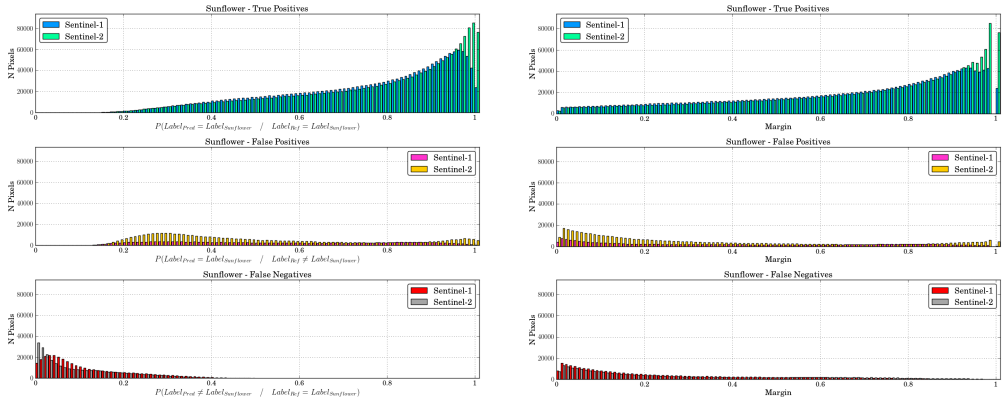
Soybean class results



(a) (b)

Figure A.11: Probability (a) and Margins (b) histograms for the Soybean class.

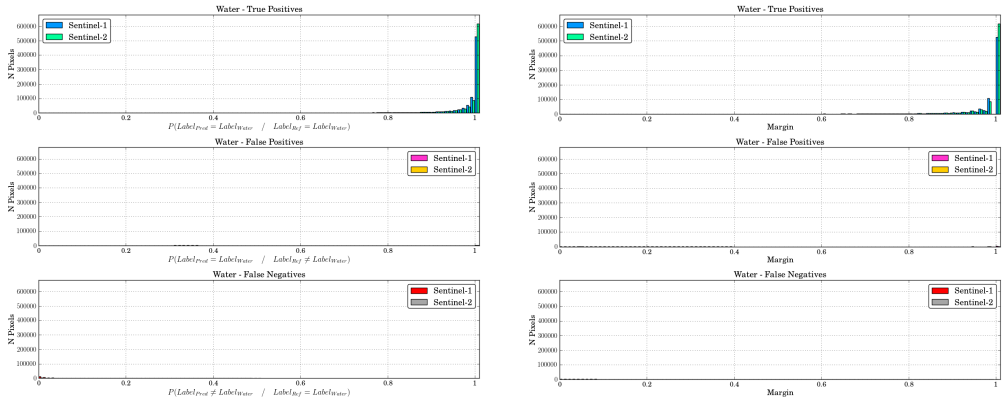
Sunflower class results



(a) (b)

Figure A.12: Probability (a) and Margins (b) histograms for the Sunflower class.

Water class results



(a) (b)

Figure A.13: Probability (a) and Margins (b) histograms for the Water class.

Appendix B

Evaluation of the fusion strategies

This appendix present the different evaluations performed for the fusion strategies. Firstly, the precision, recall and F-Score metrics are given for each class. Secondly, the same metrics are presented for the new probabilities estimation approach presented in this work. Thirdly, the analysis of the confusions between classes is given for the S1, S2 and BB strategies. Fourthly, a statistical evaluation of the classification agreements is presented for the different fusion methods. Fifthly, a visual evaluation is given presenting a set of maps that shows the classification agreements. Lastly, a summary of the operation principle and the advantages and drawbacks of the fusion methods is brief detailed.

B.1 Precision, Recall and F-Score results

Straw class results

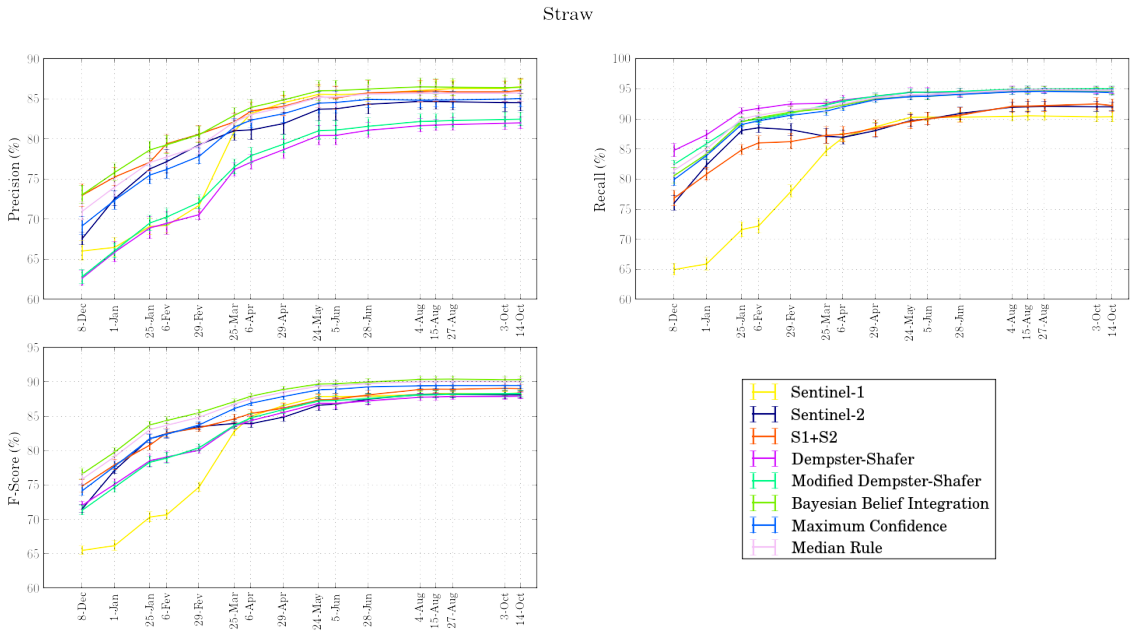


Figure B.1: Straw metrics

Comments Taking a look on Precision and Recall pictures, it could be notice the same trend as *Vine* metrics where DS based methods obtain lower precision values. But they follow a great performance, as well as the other fusion techniques, for the Recall metric. It is worth mentioning how the S1 classifier reaches outstanding rates being greater than the S2 classifier at the end of the season. Thus, the fusion techniques are removing FN samples from S1 and S2 classifications but only the pure probabilistic combiners are eliminating the FP predictions. Also, thanks to the early great S2 performance, the fusion methods are able to achieve, quickly, excellent rates for the F-score figure.

Maize class results

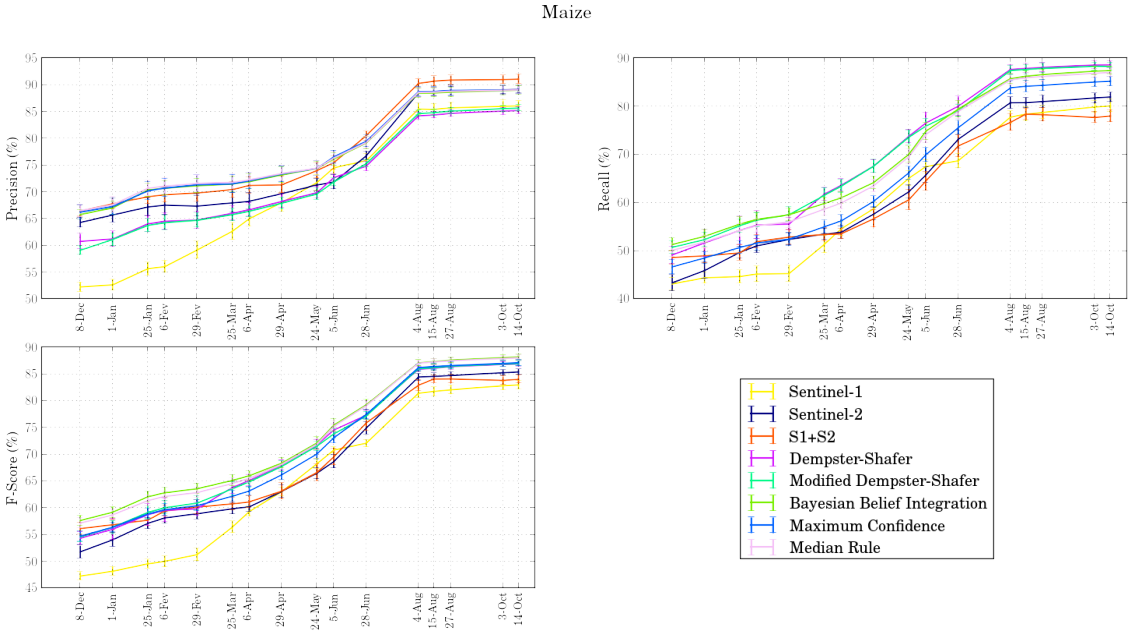


Figure B.2: Maize metrics

Comments *Maize* class presents a great performance. Nonetheless, these significant rates are only achieved beyond the spring end this is due to the fact that this class is a summer crop. In terms of precision, S2 classifiers outperforms S1 strategy showing a better performance because of the lower number of FP predictions. For the *Maize* case all the combination techniques present greater results than the single classifiers.

Soybean class results

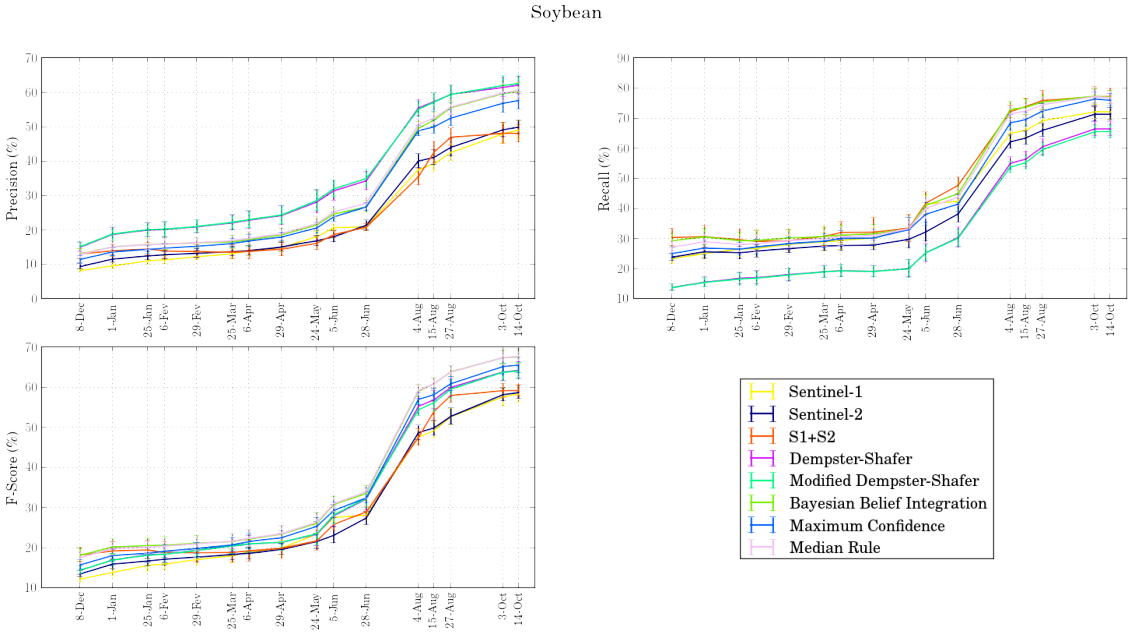


Figure B.3: Soybean metrics

Comments *Soybean* class shows an limited performance. Unlike the previous classes, for this crop, Dempster-Shafer based methods reach the highest rate for the Precision metric. This class presents an important confusions with *Maize* class. *Soybean* is a summer crop like *Sorghum* or *Maize* classes. But, neither the performance nor the confusion are not worse than *Sorghum* results. It seems that the probabilistic combiners are able to remove either FP or FN predictions.

Alfalfa class results

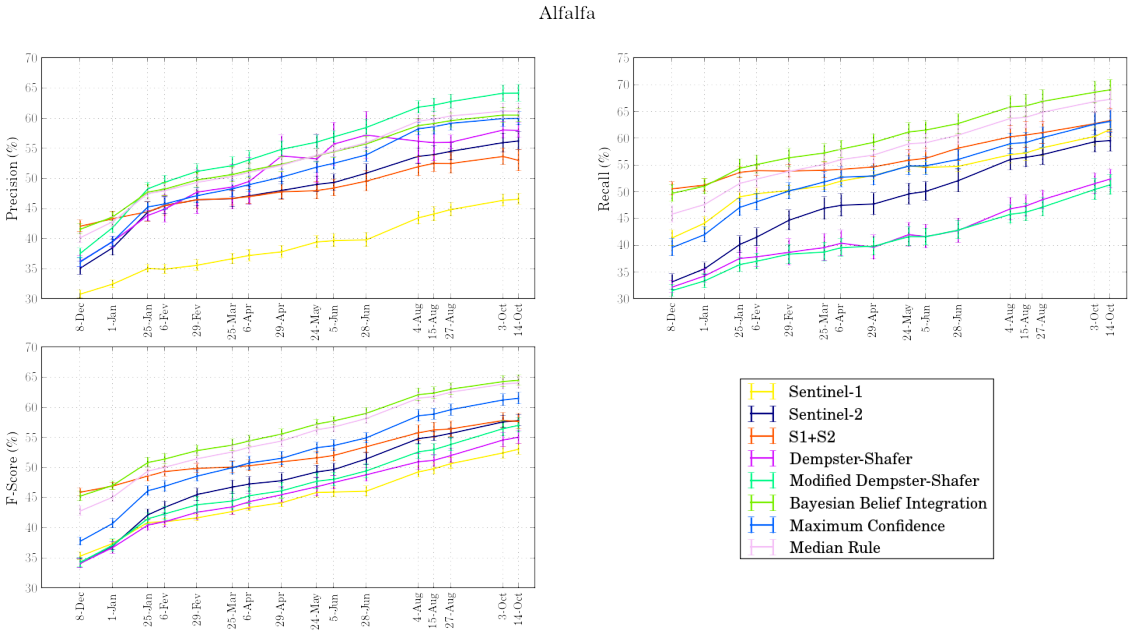


Figure B.4: Alfalfa metrics

Comments *Alfalfa* crop meets similar characteristics than *Sorghum* and *Soybean* classes which is why it follows a similar performance. In this case, the confusions matrices, from the the single classifications, present a large number of FP related to *Grassland* class. F-Score results show that the pure probabilistic techniques obtain an important improvement of the performance.

Grassland class results

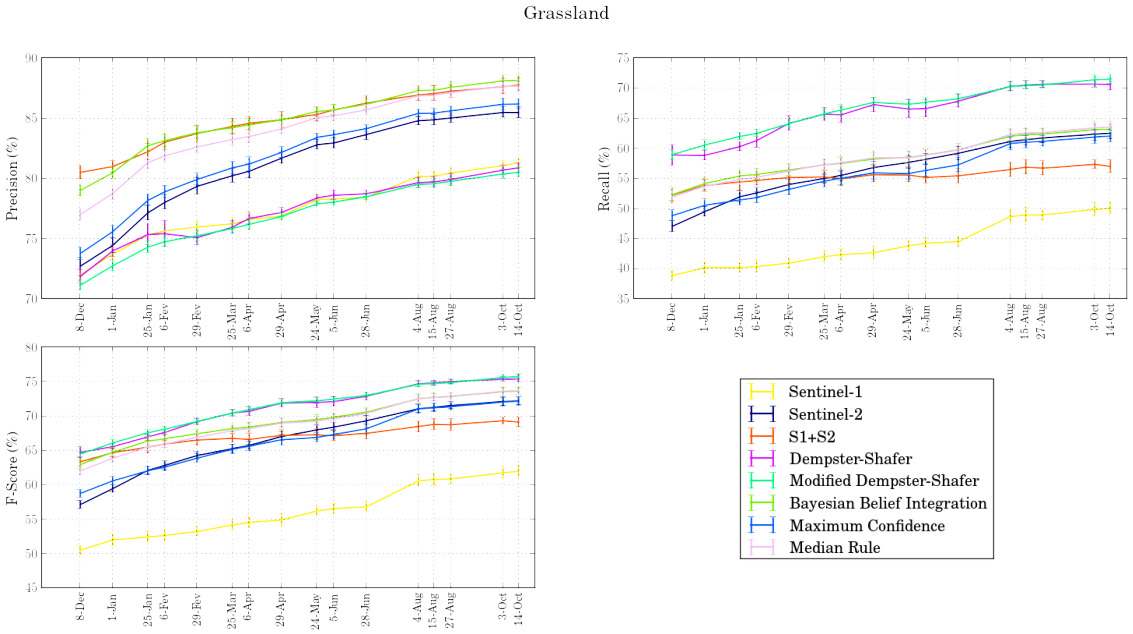


Figure B.5: Grassland metrics

Comments Despite the poor S1 strategy performance because of the considerable confusion between *Fallow* and *Alfalfa* classes, fusion techniques achieve significant results for this class. Specially, Dempster-Shafer based methods are able to enhance the number of FN samples and for the Recall and F-Score metrics, outperform any other strategy.

Fallow class results

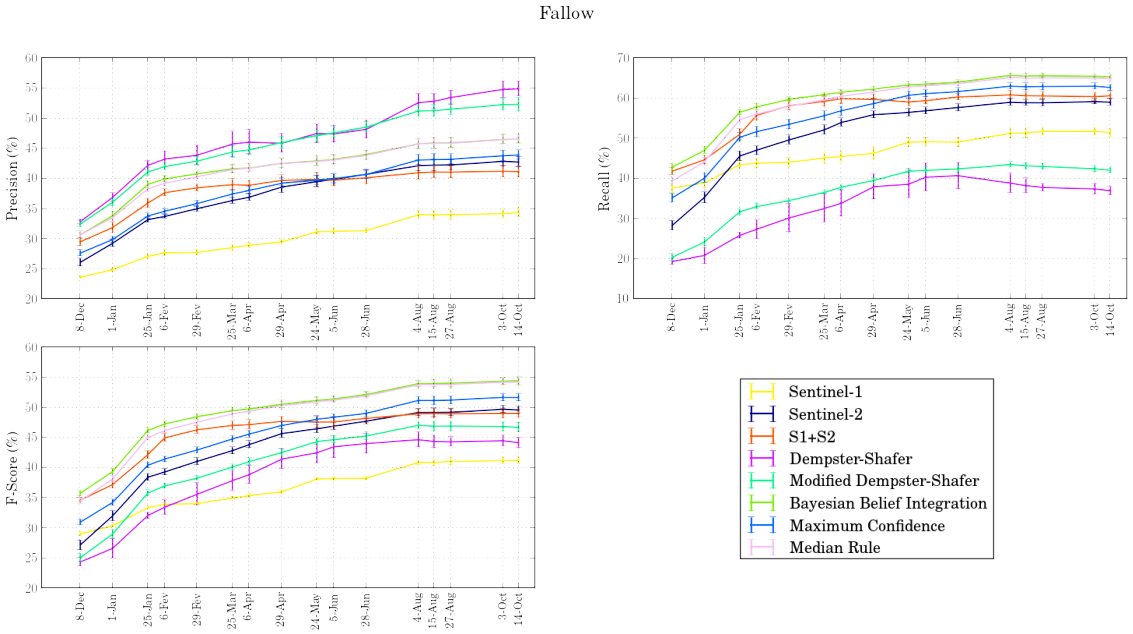


Figure B.6: Fallow metrics

Comments In the same way as *Grassland* class, *Fallow* class classification experiences a considerable amount of confusions between them being the FP samples the most widespread. Pure probabilistic combiners present a better performance achieving the best metric results. Dempster-Shafer based methods show excellent Precision rates but, in contrast, the lowest performance for the Recall resulting in an inefficient way to fuse the single classifiers as F-Score metric shows.

Shrubland class results

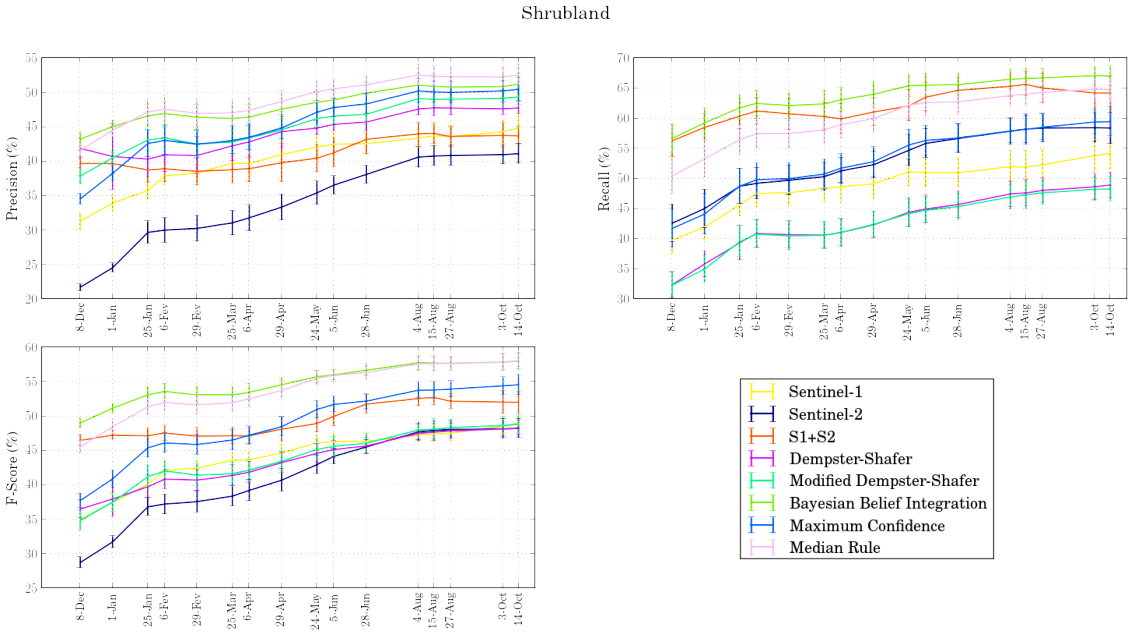


Figure B.7: Shrubland metrics

Comments *Shrubland* class presents the similar performance for both single classifiers. Fusion methods carry an important improvement, specially the BB strategy and the MR. Both obtain the best results for each metric. In contrast, Dempster-Shafer based methods are able to improve S1 and S2 classifiers for the precision rates but not for the recall where they obtain the lowest scores resulting in a inefficient fusion as F-Score metric indicates.

Rapeseed class results

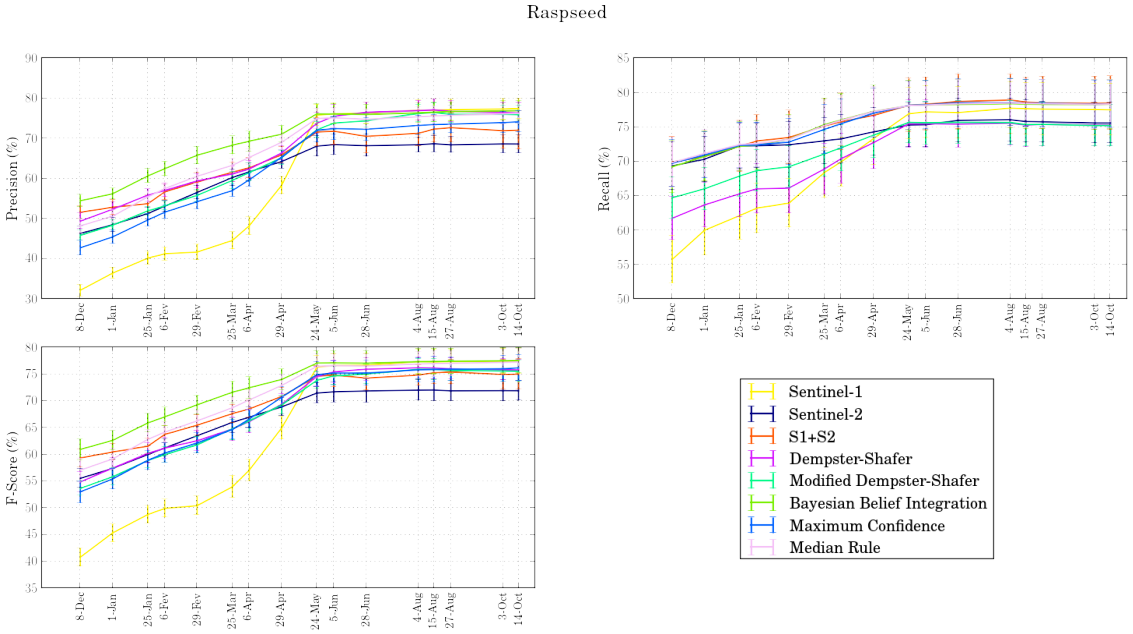


Figure B.8: Raspseed metrics

Comments Rapeseed crops are characterized by obtain a great performance for the S1 classifier, outperforming the S2 strategy, but increasing its efficiency at the end of the spring. Nonetheless, it is a class where the presented fusion methods do not imply an extraordinary enhancement at the end of the season. But, they are able to achieve excellent rates from the very beginning, unlike S1 and S2 classifiers.

Deciduous

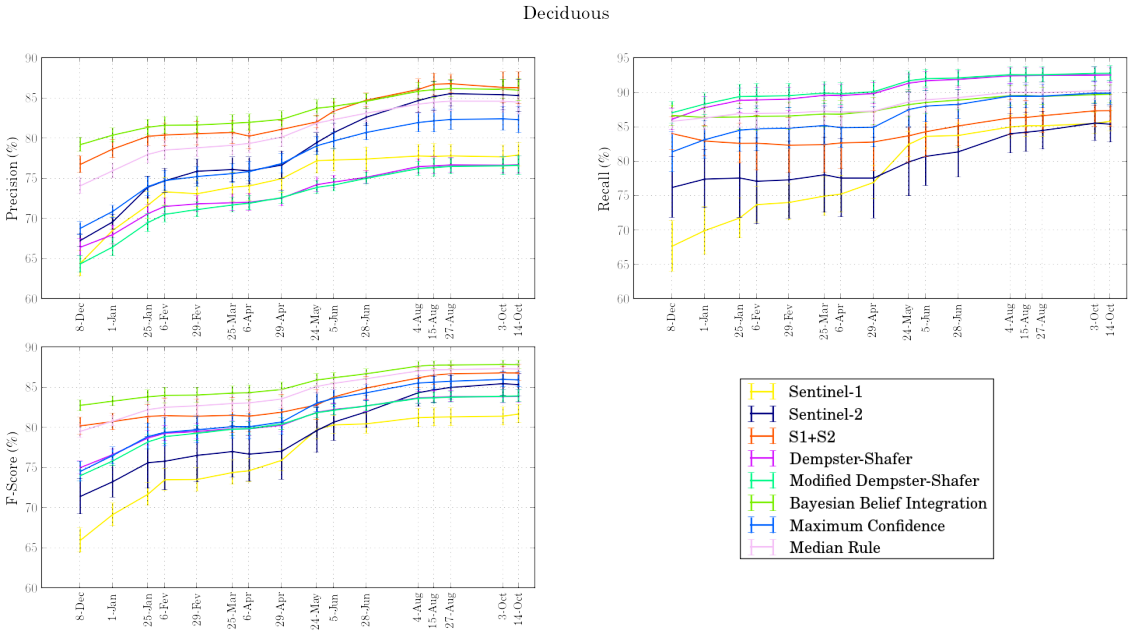


Figure B.9: Deciduous metrics

Comments Deciduous is a permanent class which already achieves important results without the fusion step. Also, fusion techniques as BB strategy or the MR obtain some great results along all the agricultural season. As the Recall figure shows, Dempster-Shafer methods presents the best capability of removing FN predicted samples. But, in contrast, for the Precision metric they display the worst results. It is worth highlight how the S2 classifier by itself is able to overcome the most of the solutions for the Precision plot but obtaining the worst performance for the Recall. Also, for the F-Score metric it is only outperformed by the probabilistic fusion methods.

Water

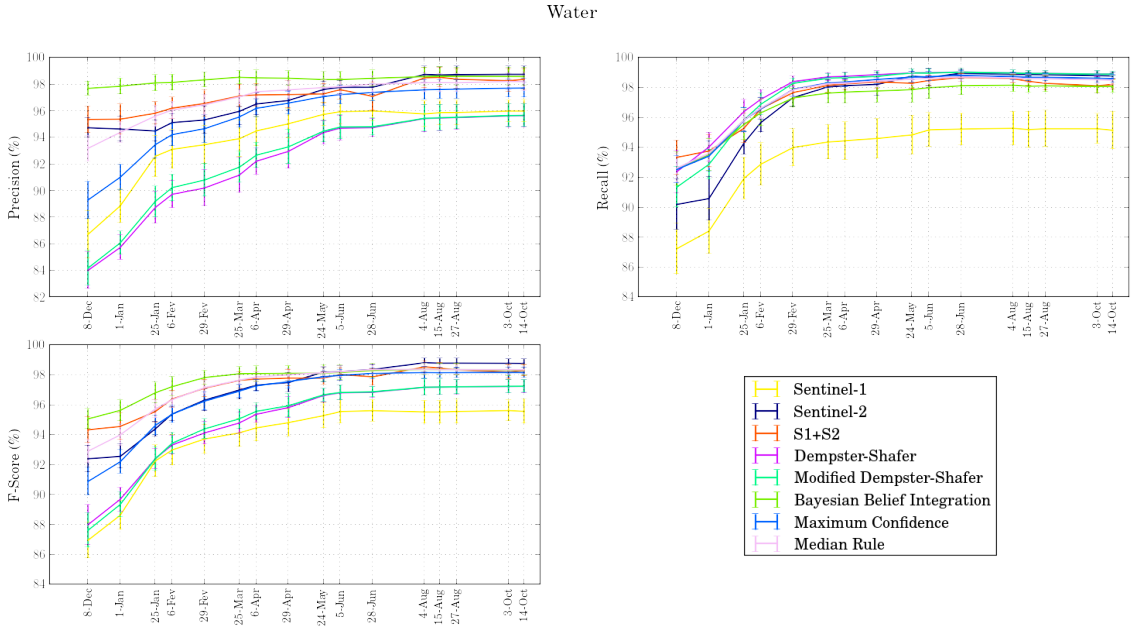


Figure B.10: Water metrics

Comments As a permanent class, its particular spectral radiance and surface leads water to the best classification results provided by the S2 classifier. The fusion techniques improve the performance in terms of time, specially the probabilistic combiners which are able to achieve excellent results from the very beginnings of the season.

Orchard

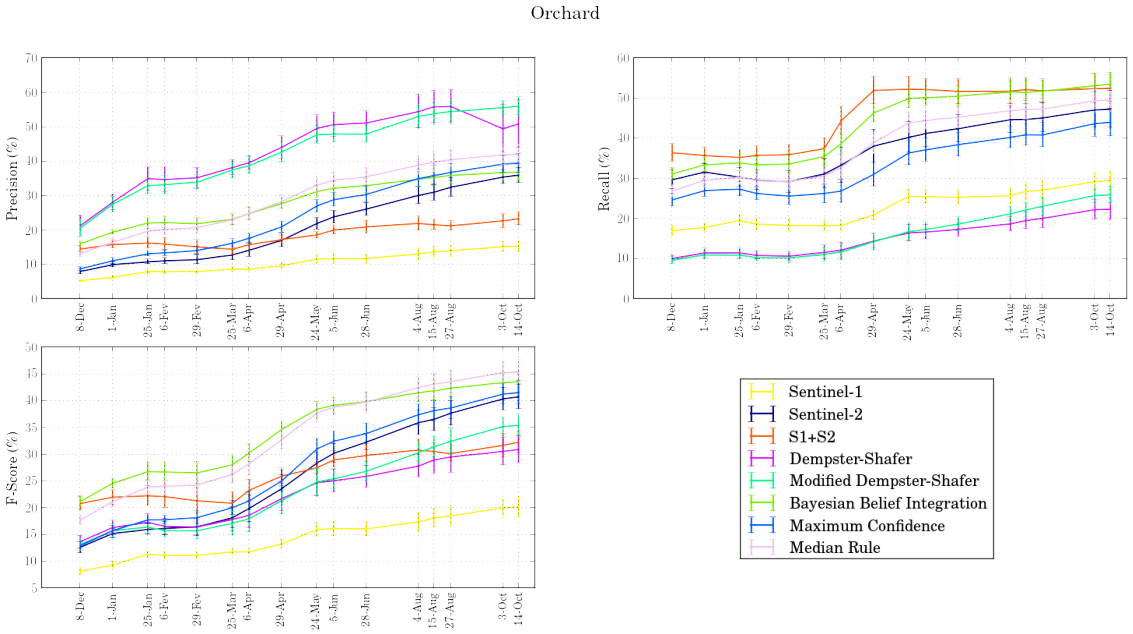


Figure B.11: Orchard metrics

Comments In contrast with *Water* class, *Orchard* class obtains the poorest performance. One of the reason could be the strong confusion between *Vine* and *Grassland* classes that the models present. Nonetheless, pure probabilistic fusion methods achieve greater results in comparison with the S1 and S2 classifiers and the Dempster-Shafer based methods.

B.2 Precision, Recall and F-Scores results for the new class probabilities estimation approach

Alfalfa class results

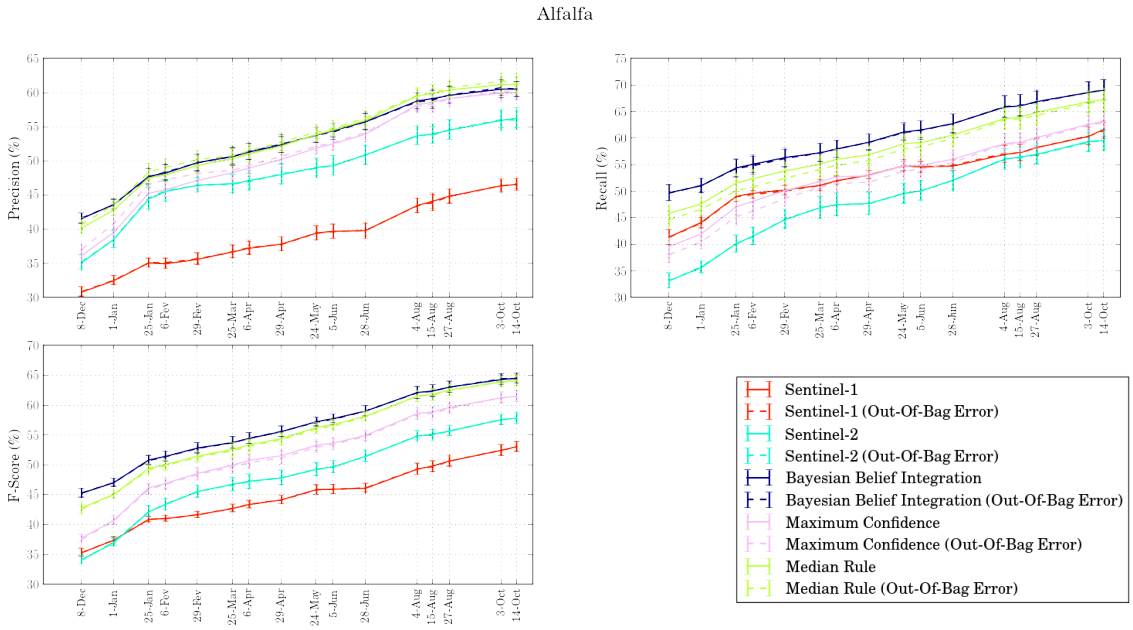


Figure B.12: Alfalfa metrics results for the new probabilities estimation approach.

Build up class results

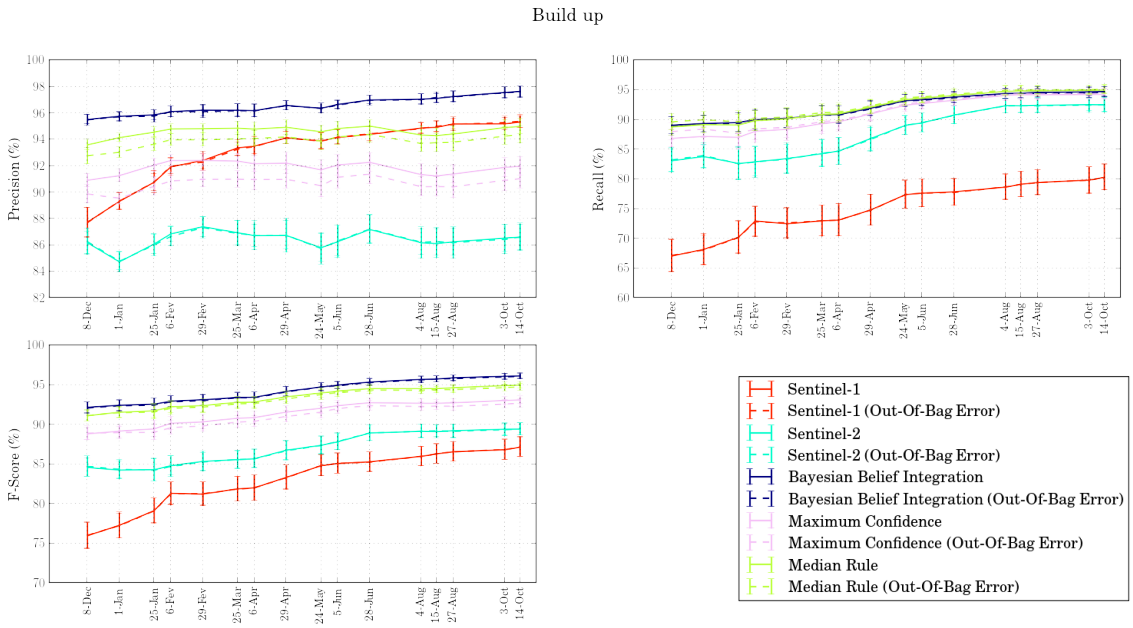


Figure B.13: Build up metrics results for the new probabilities estimation approach.

Deciduous class results

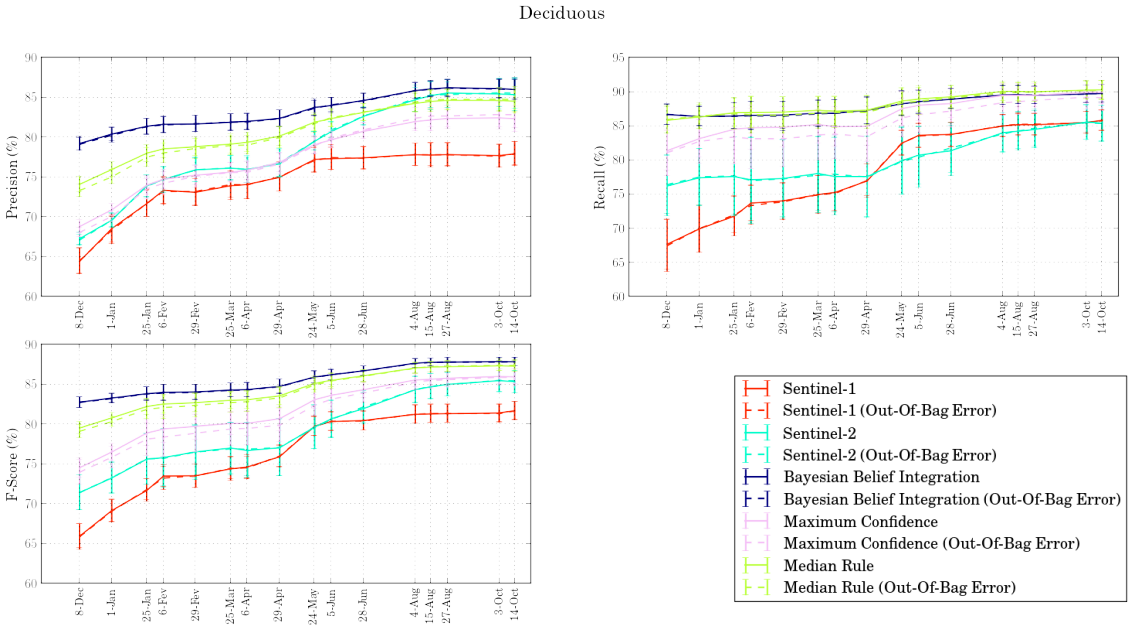


Figure B.14: Deciduous metrics results for the new probabilities estimation approach.

Fallow class results

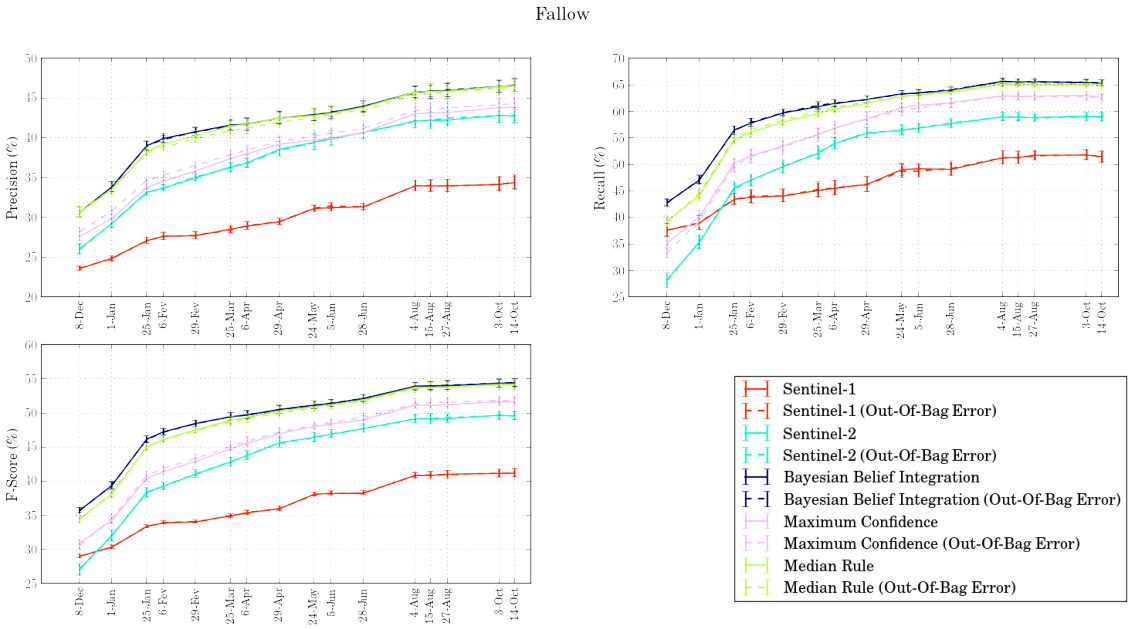


Figure B.15: Fallow metrics results for the new probabilities estimation approach.

Grassland class results

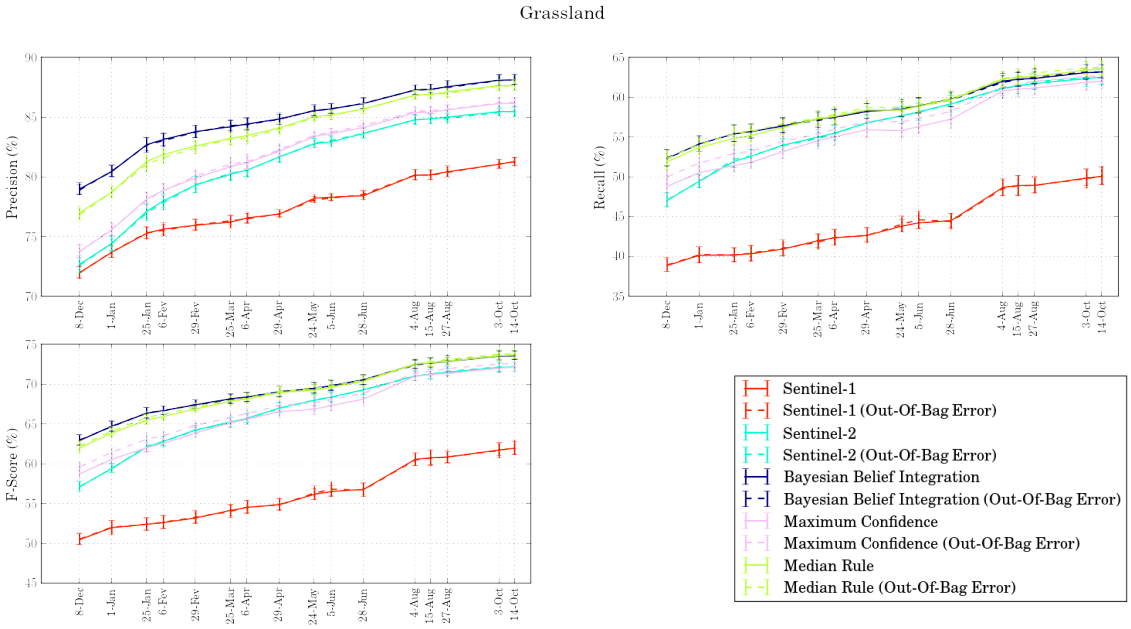


Figure B.16: Grassland metrics results for the new probabilities estimation approach.

Maize class results

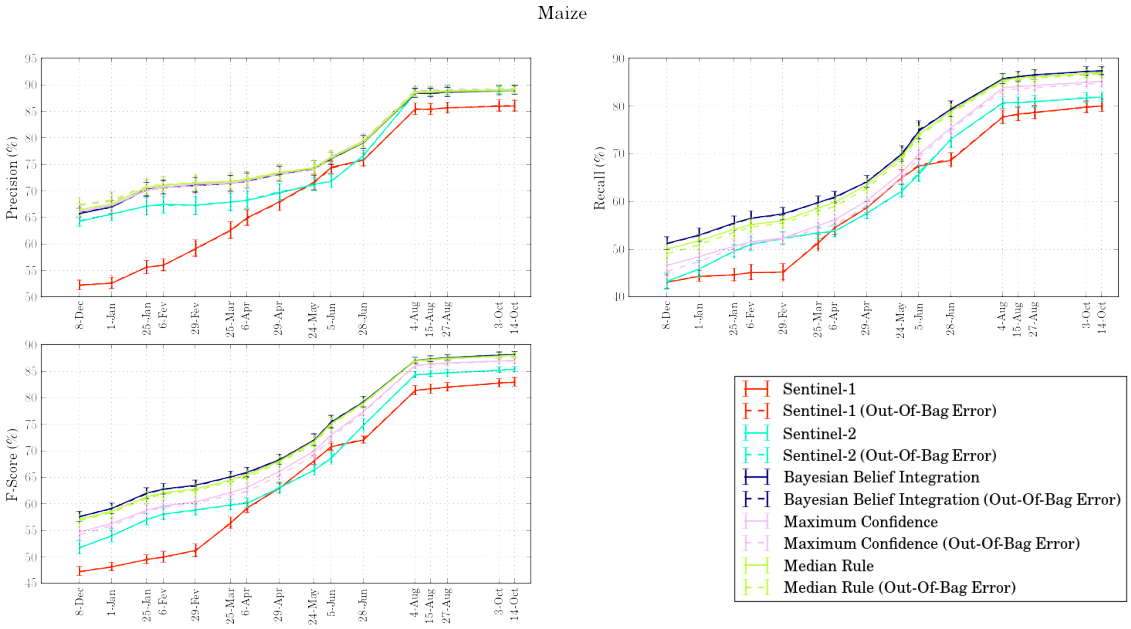


Figure B.17: Maize metrics results for the new probabilities estimation approach.

Orchard class results

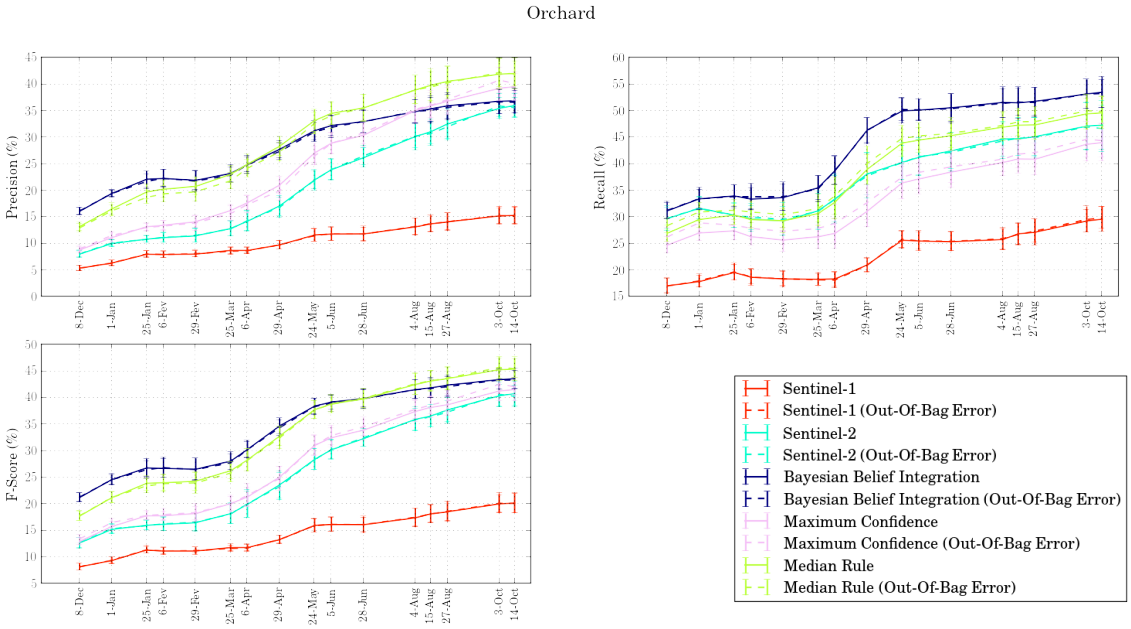


Figure B.18: Orchard metrics results for the new probabilities estimation approach.

Rapeseed class results

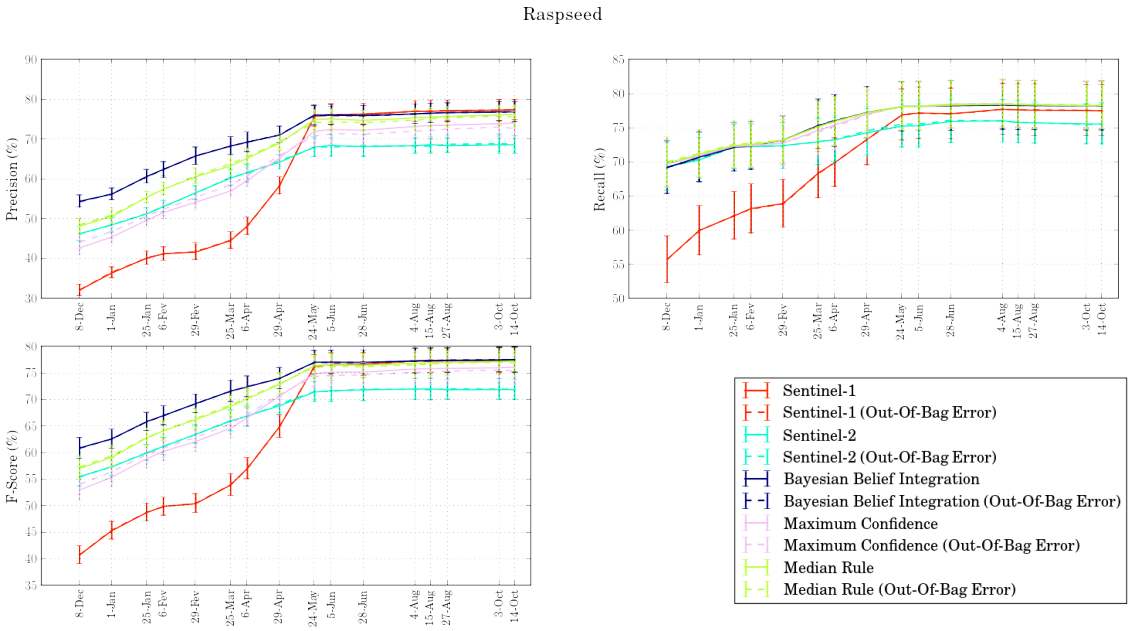


Figure B.19: Rapeseed metrics results for the new probabilities estimation approach.

Sorghum class results

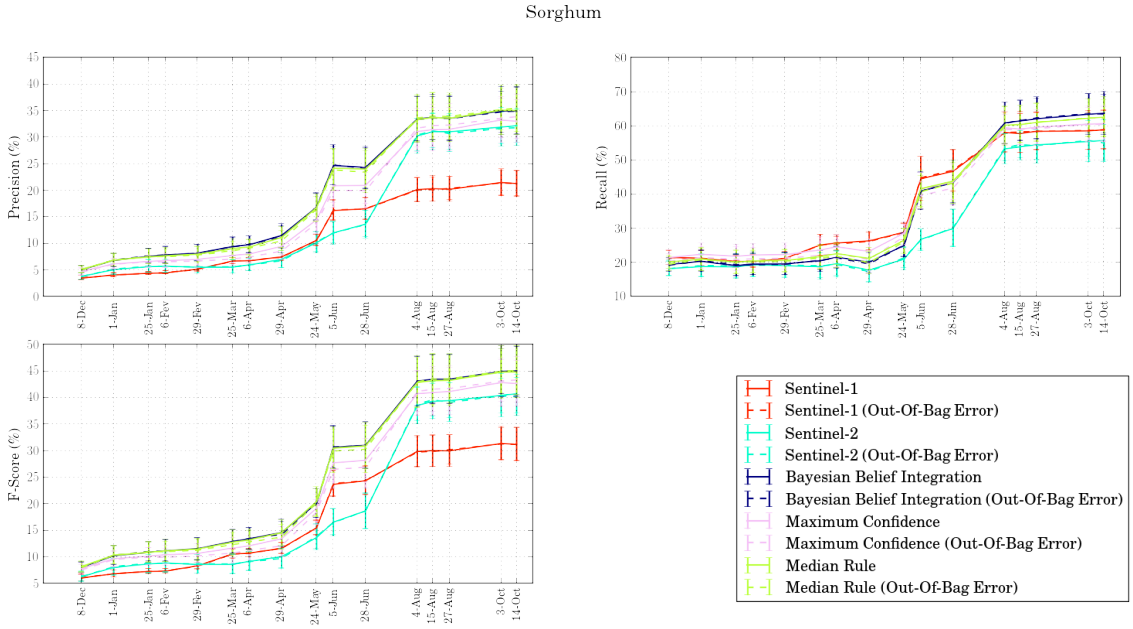


Figure B.20: Sorghum metrics results for the new probabilities estimation approach.

Soybean class results

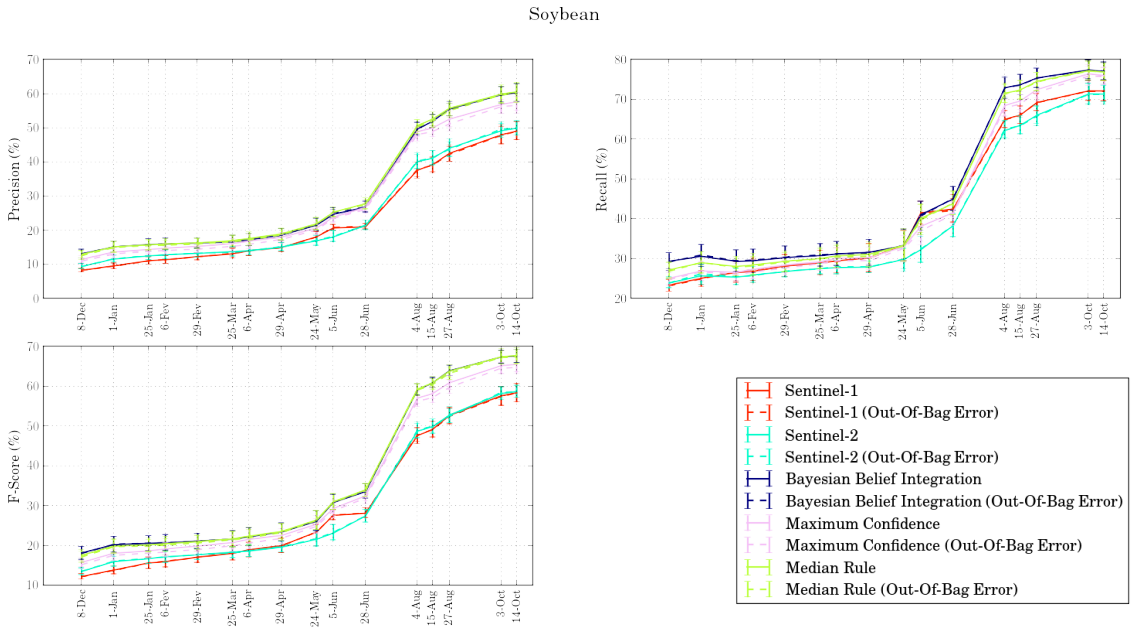


Figure B.21: Soybean metrics results for the new probabilities estimation approach.

Straw class results

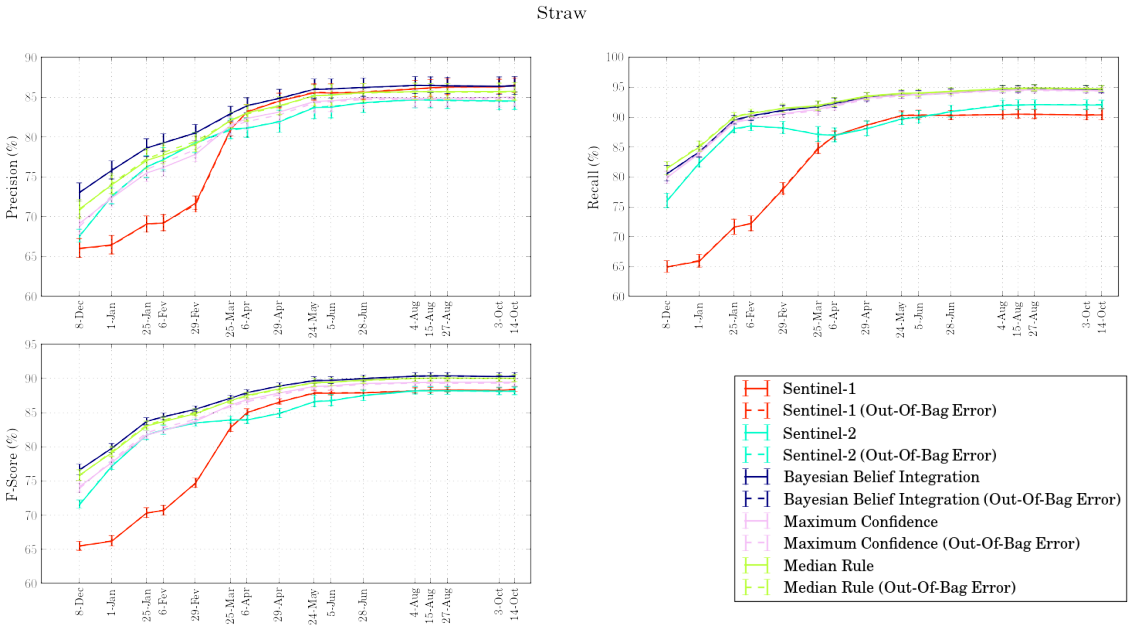


Figure B.22: Straw metrics results for the new probabilities estimation approach.

Sunflower class results

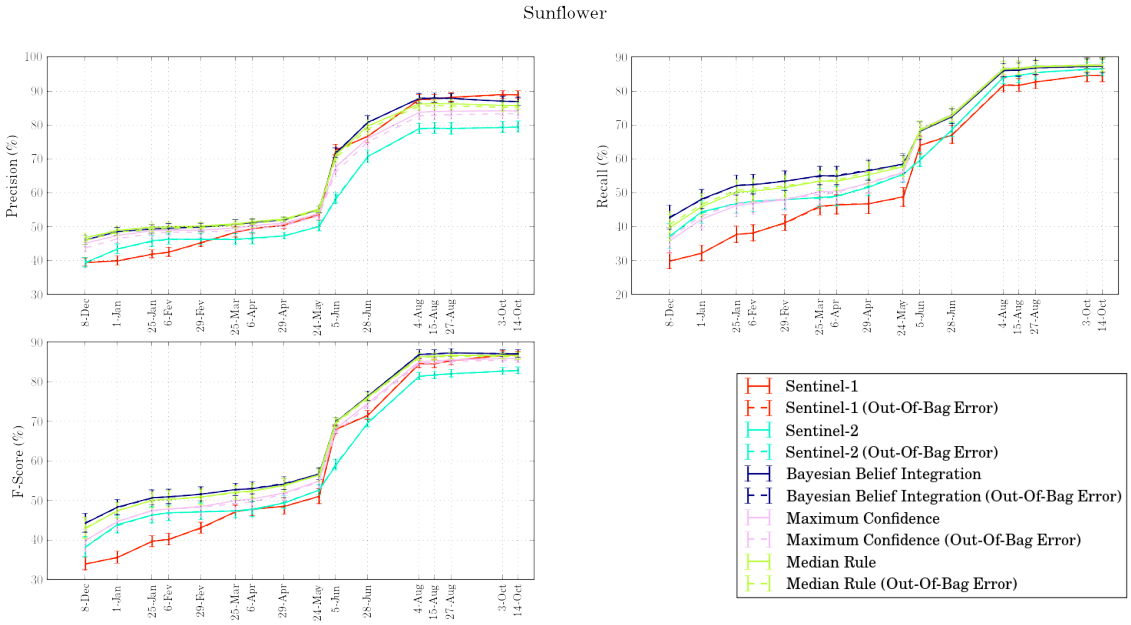


Figure B.23: Sunflower metrics results for the new probabilities estimation approach.

Vine class results

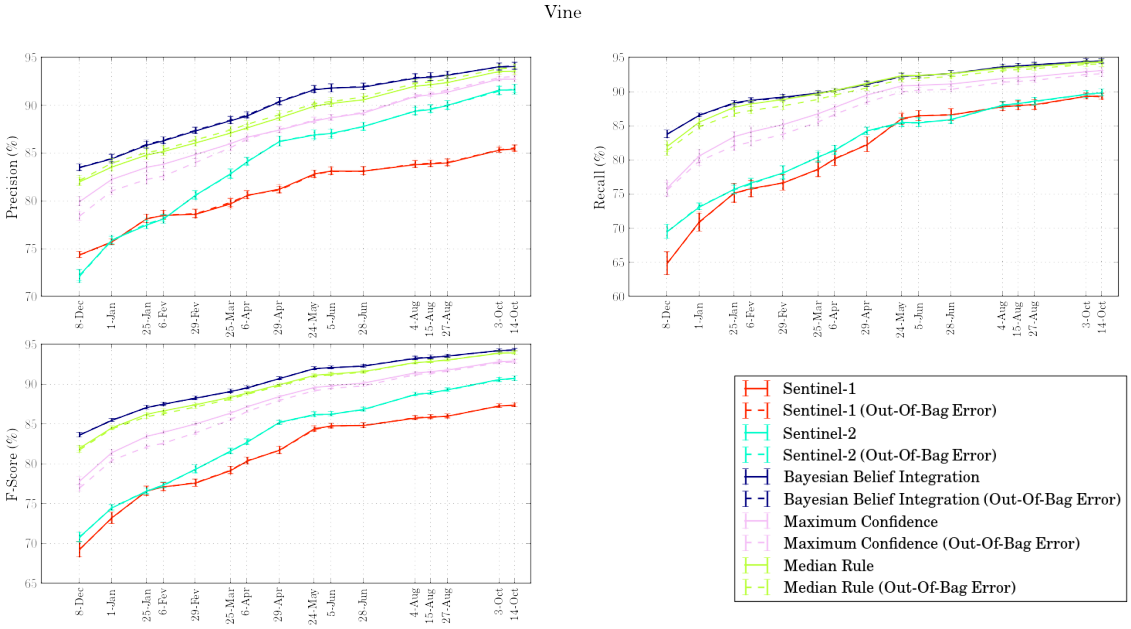


Figure B.24: Vine metrics results for the new probabilities estimation approach.

Water class results

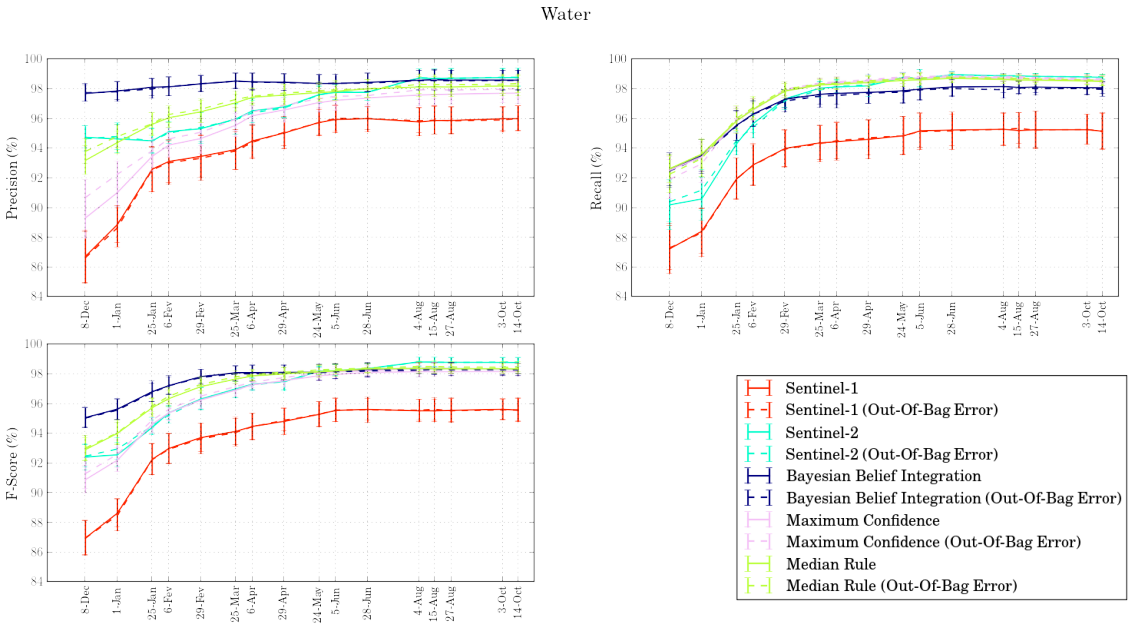
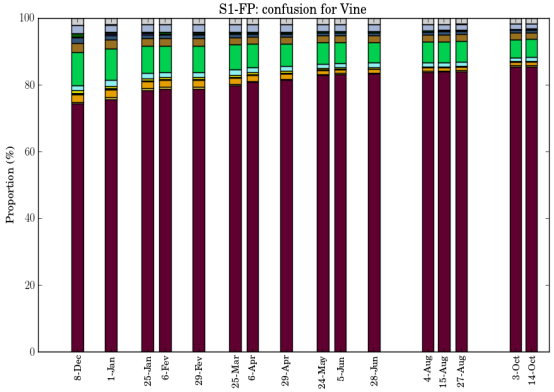
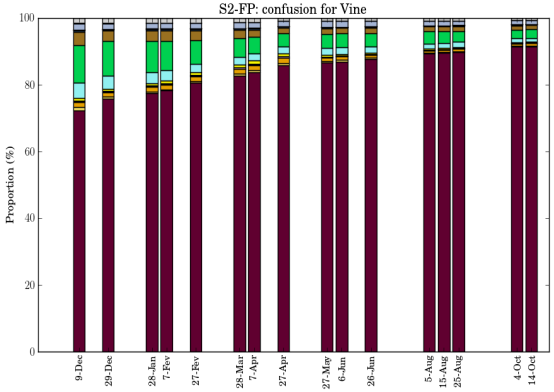


Figure B.25: Water metrics results for the new probabilities estimation approach.

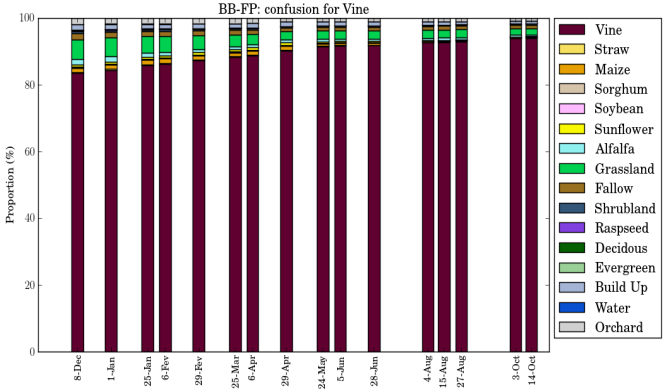
B.3 Analysis of the confusions between classes



(a) Sentinel-1

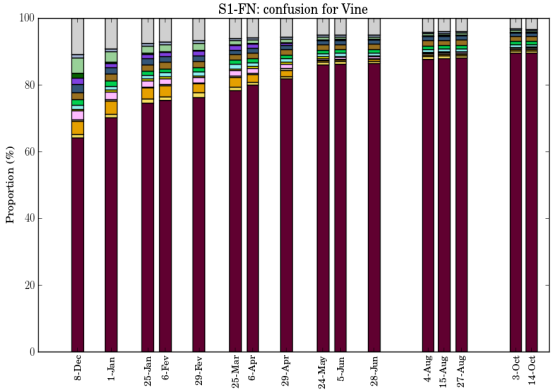


(b) Sentinel-2

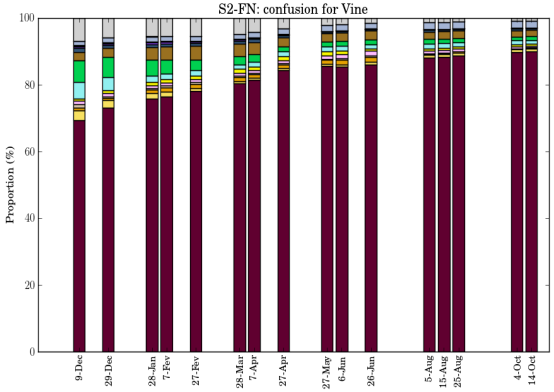


(c) Bayesian Belief Integration

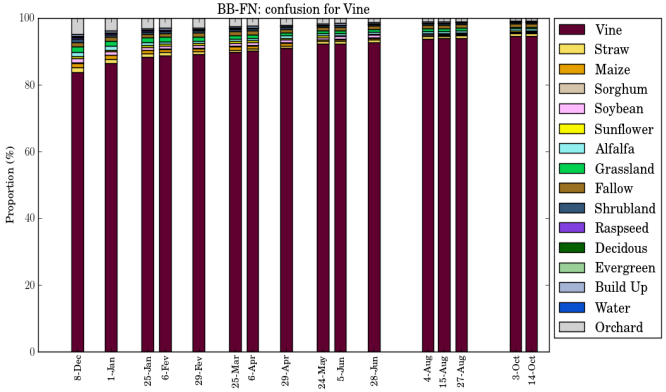
Figure B.26: FP confusion approach for the *Vine* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.



(a) Sentinel-1

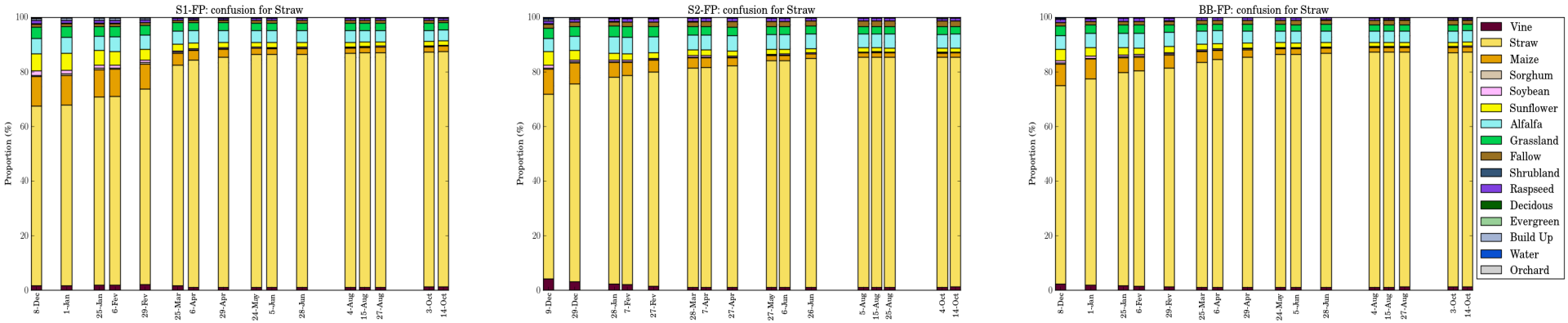


(b) Sentinel-2



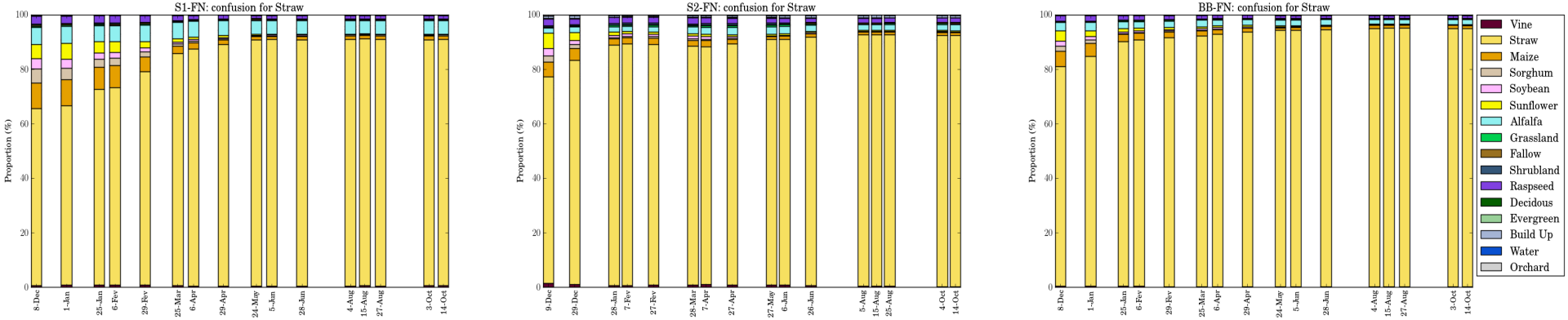
(c) Bayesian Belief Integration

Figure B.27: FN confusion approach for the *Vine* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.



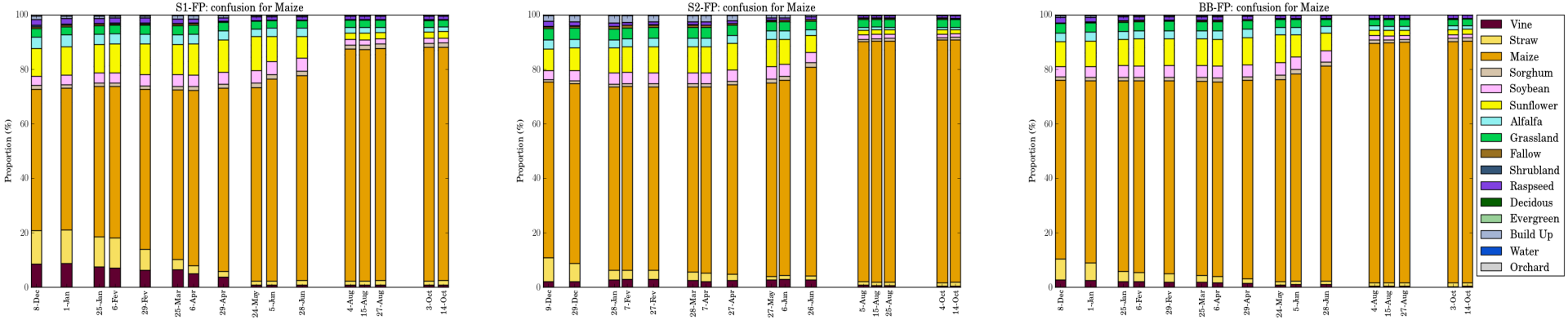
(a) Sentinel-1 (b) Sentinel-2 (c) Bayesian Belief Integration

Figure B.28: FP confusion approach for the *Straw* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.



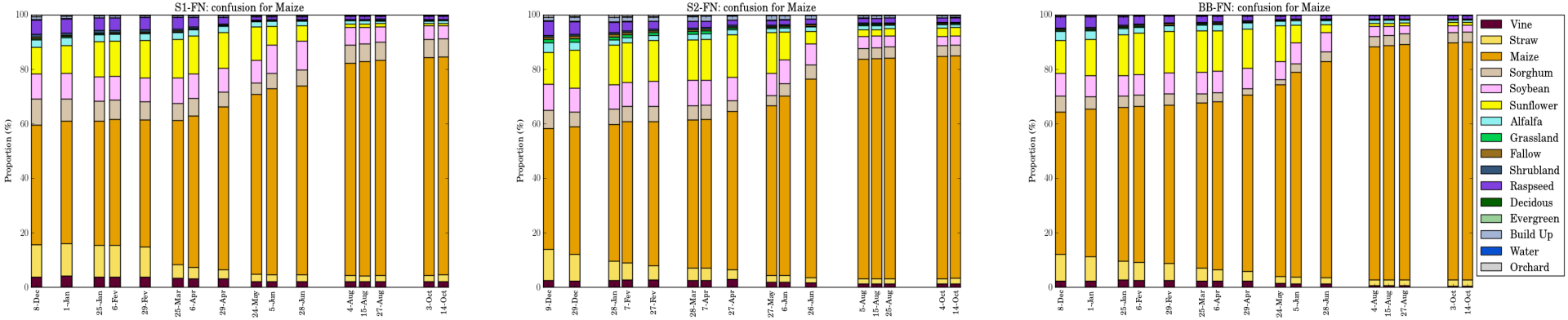
(a) Sentinel-1 (b) Sentinel-2 (c) Bayesian Belief Integration

Figure B.29: FN confusion approach for the *Straw* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.



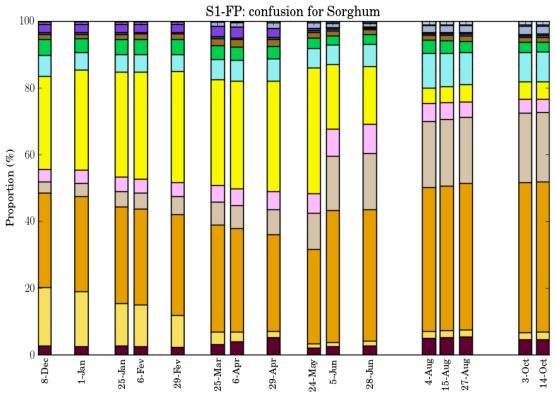
(a) Sentinel-1 (b) Sentinel-2 (c) Bayesian Belief Integration

Figure B.30: FP confusion approach for the *Maize* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.

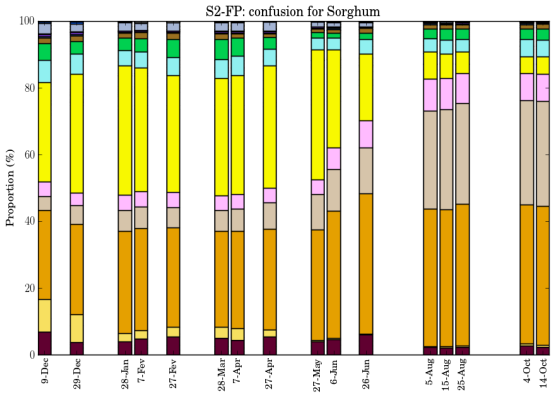


(a) Sentinel-1 (b) Sentinel-2 (c) Bayesian Belief Integration

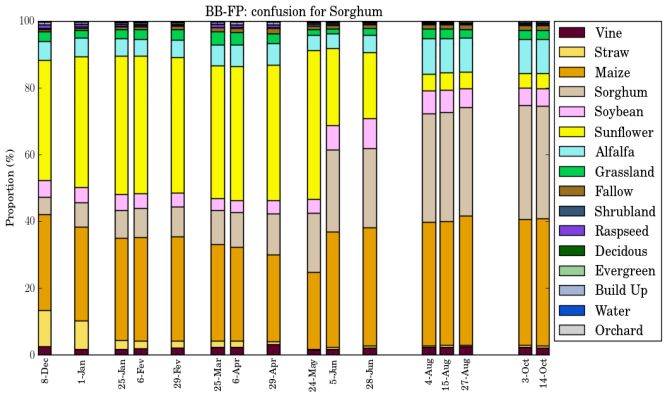
Figure B.31: FN confusion approach for the *Maize* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.



(a) Sentinel-1

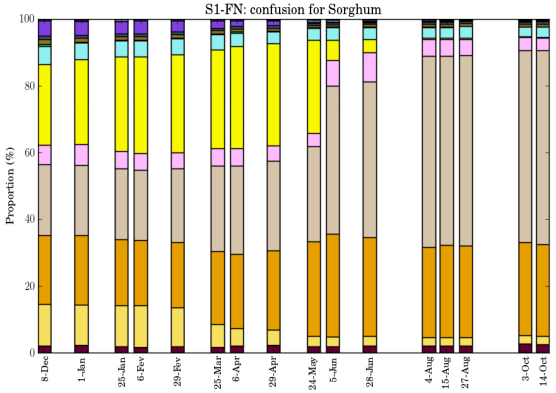


(b) Sentinel-2

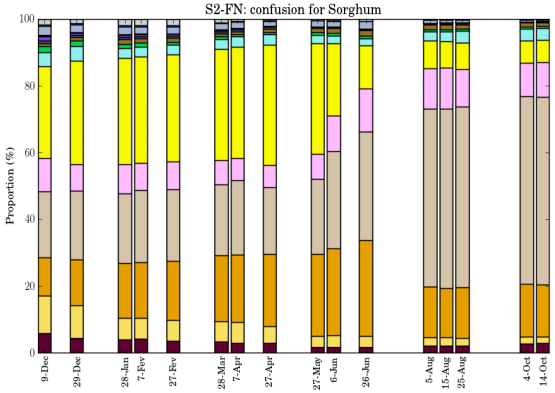


(c) Bayesian Belief Integration

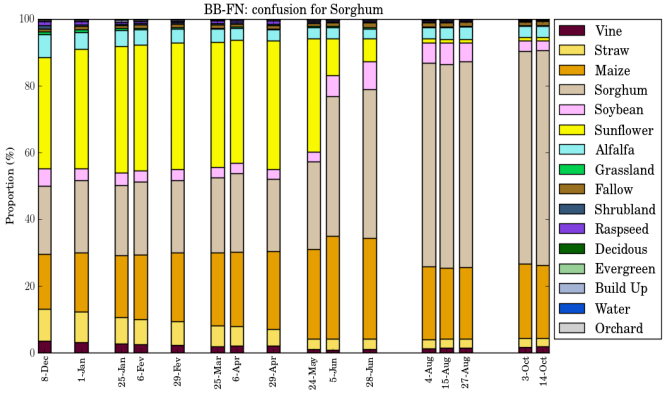
Figure B.32: FP confusion approach for the *Sorghum* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.



(a) Sentinel-1

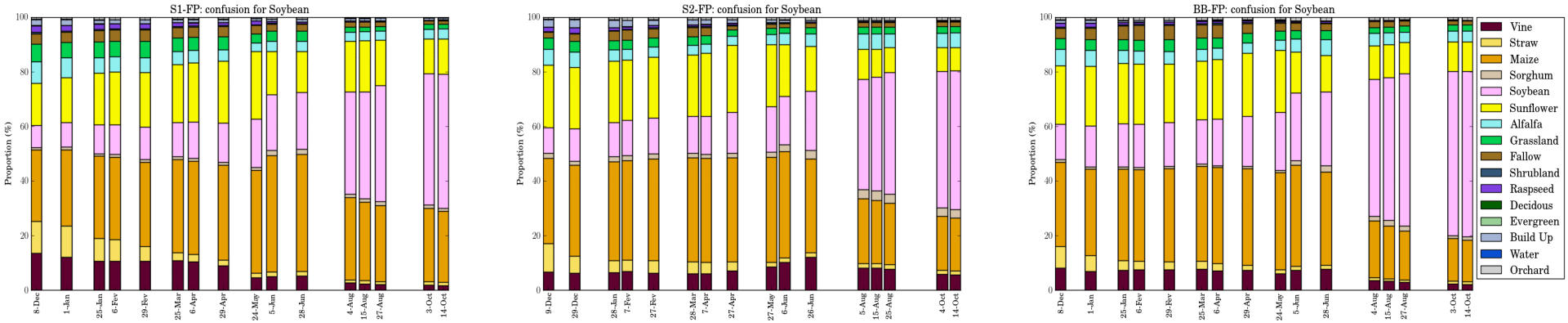


(b) Sentinel-2



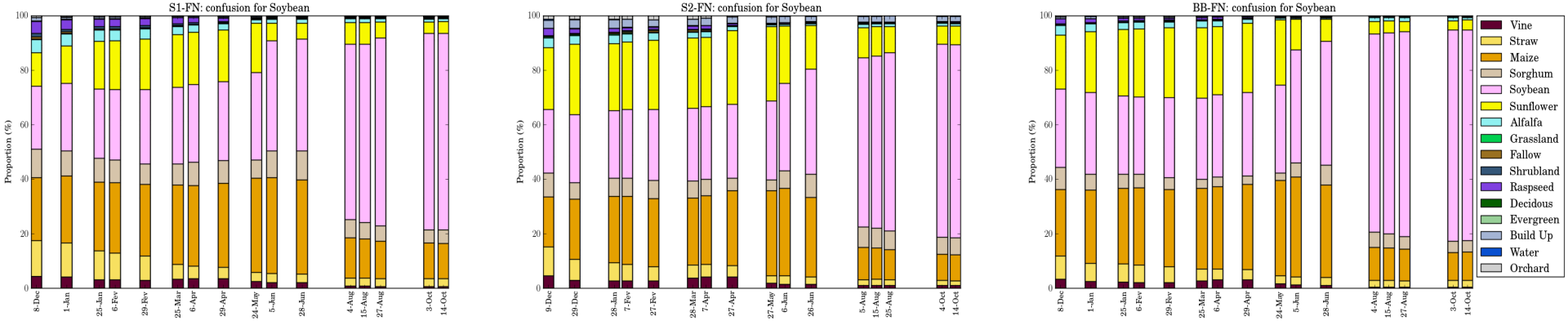
(c) Bayesian Belief Integration

Figure B.33: FN confusion approach for the *Sorghum* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.



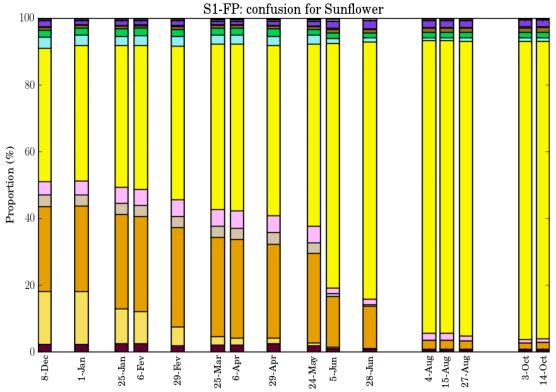
(a) Sentinel-1 (b) Sentinel-2 (c) Bayesian Belief Integration

Figure B.34: FP confusion approach for the *Soybean* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.

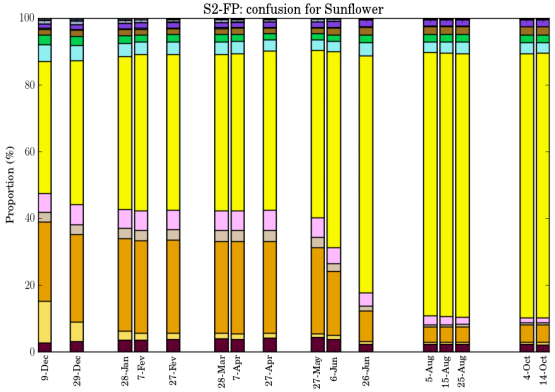


(a) Sentinel-1 (b) Sentinel-2 (c) Bayesian Belief Integration

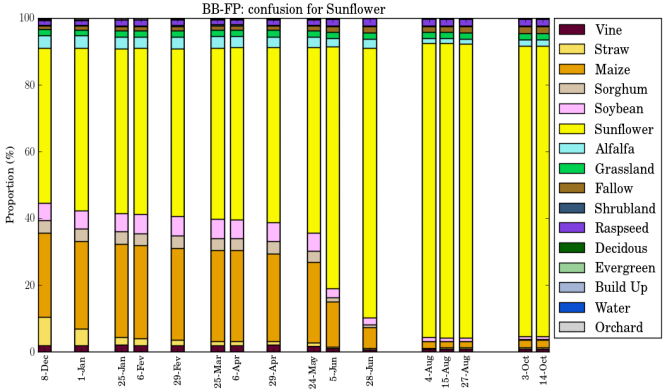
Figure B.35: FN confusion approach for the *Soybean* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.



(a) Sentinel-1

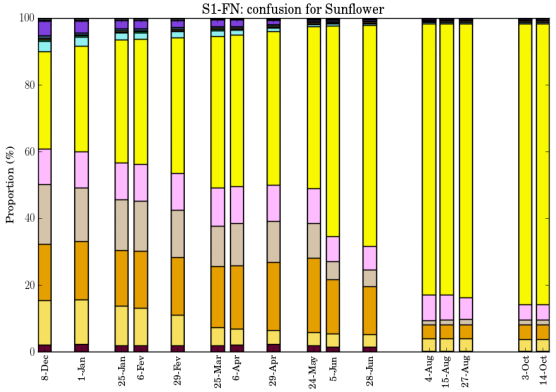


(b) Sentinel-2

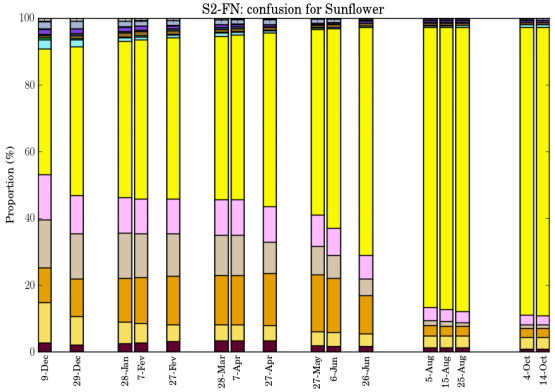


(c) Bayesian Belief Integration

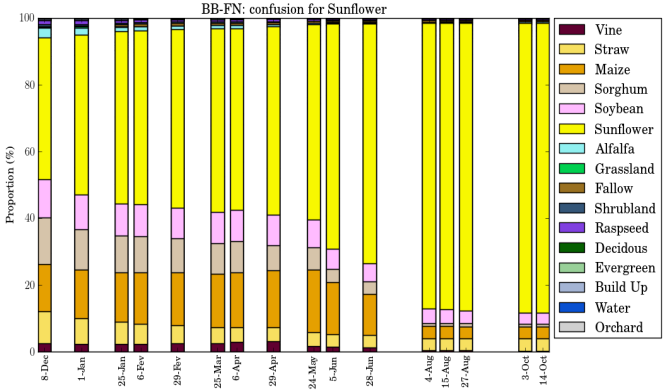
Figure B.36: FP confusion approach for the *Sunflower* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.



(a) Sentinel-1

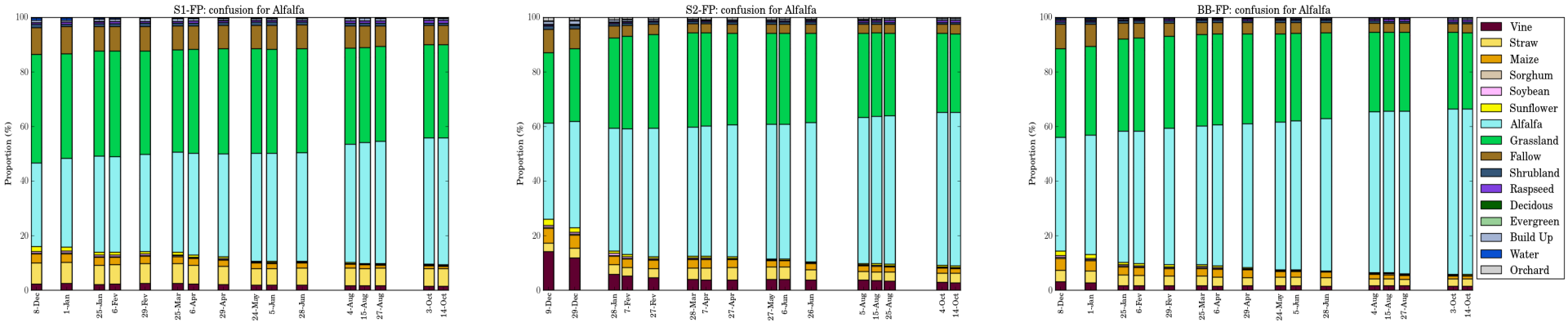


(b) Sentinel-2



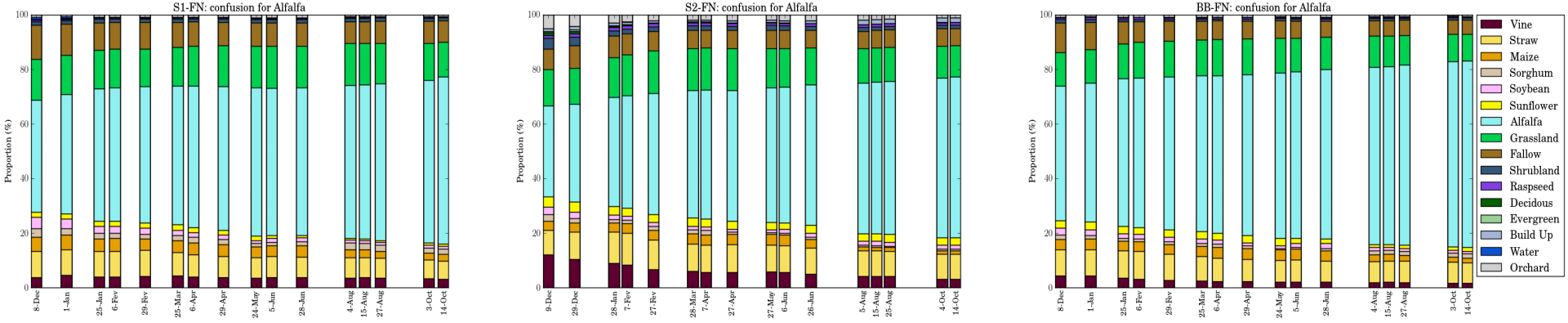
(c) Bayesian Belief Integration

Figure B.37: FN confusion approach for the *Sunflower* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.



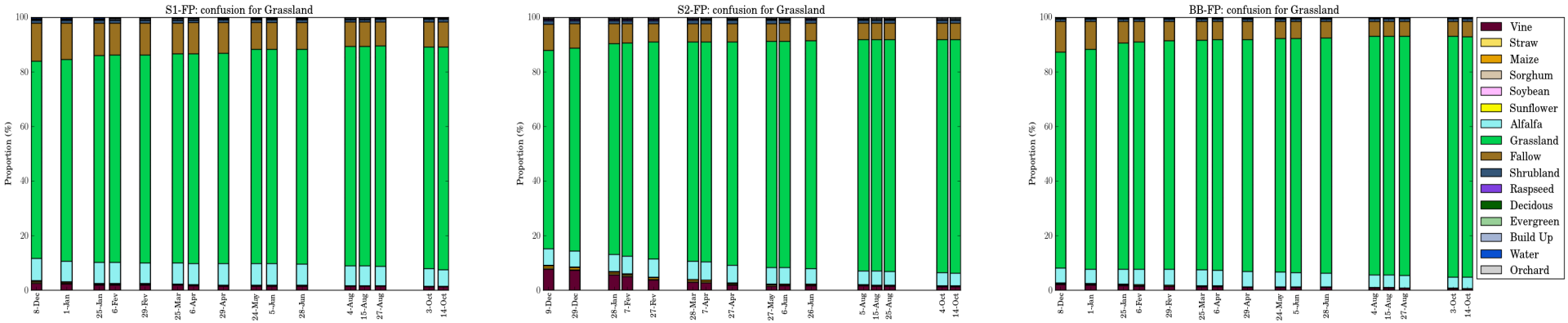
(a) Sentinel-1 (b) Sentinel-2 (c) Bayesian Belief Integration

Figure B.38: FP confusion approach for the *Alfalfa* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.



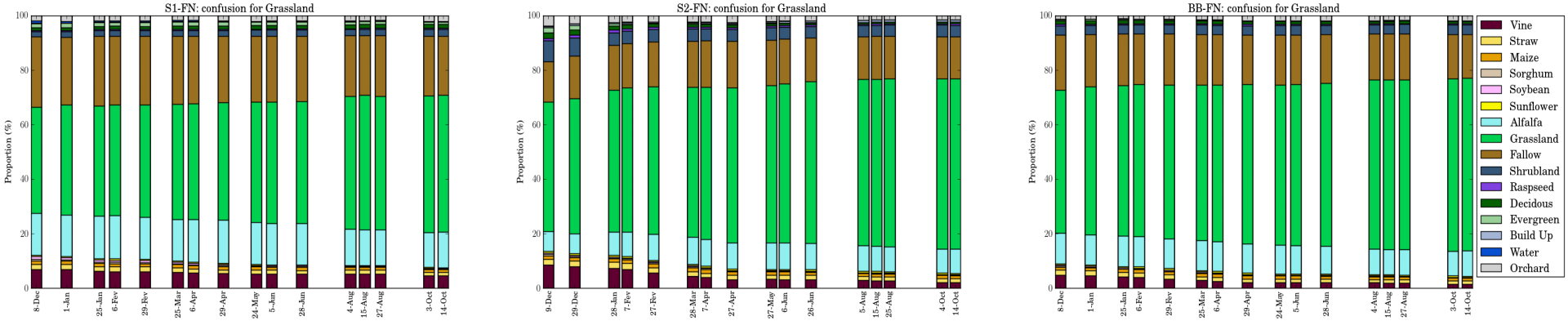
(a) Sentinel-1 (b) Sentinel-2 (c) Bayesian Belief Integration

Figure B.39: FN confusion approach for the *Alfalfa* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.



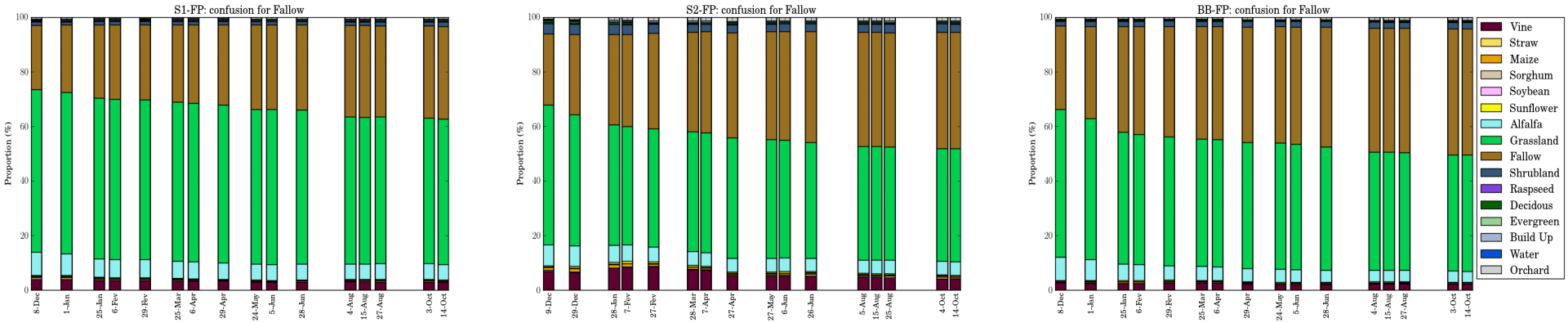
(a) Sentinel-1 (b) Sentinel-2 (c) Bayesian Belief Integration

Figure B.40: FP confusion approach for the *Grassland* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.



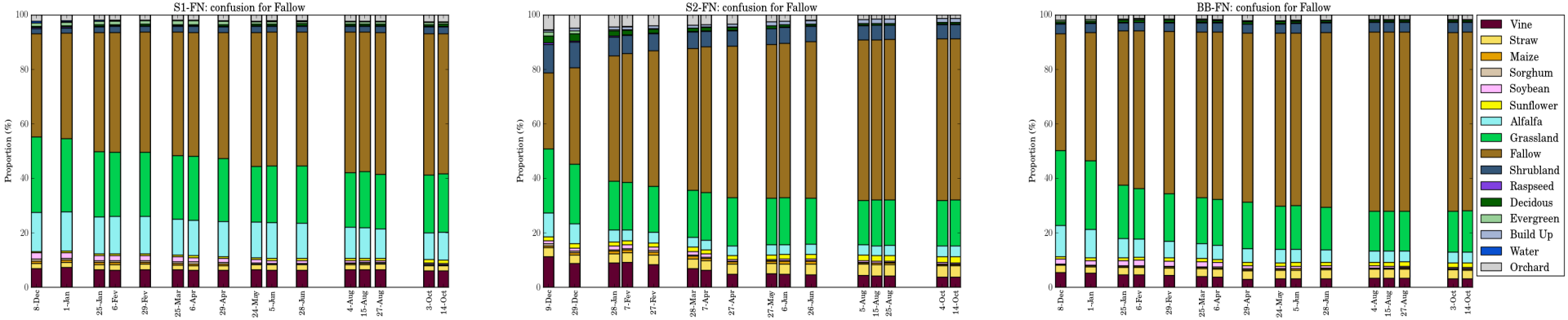
(a) Sentinel-1 (b) Sentinel-2 (c) Bayesian Belief Integration

Figure B.41: FN confusion approach for the *Grassland* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.



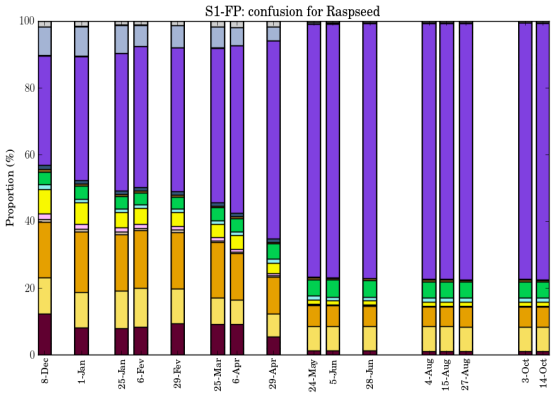
(a) Sentinel-1 (b) Sentinel-2 (c) Bayesian Belief Integration

Figure B.42: FP confusion approach for the *Fallow* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.

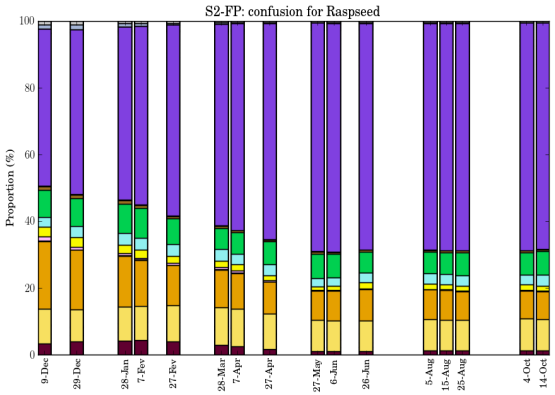


(a) Sentinel-1 (b) Sentinel-2 (c) Bayesian Belief Integration

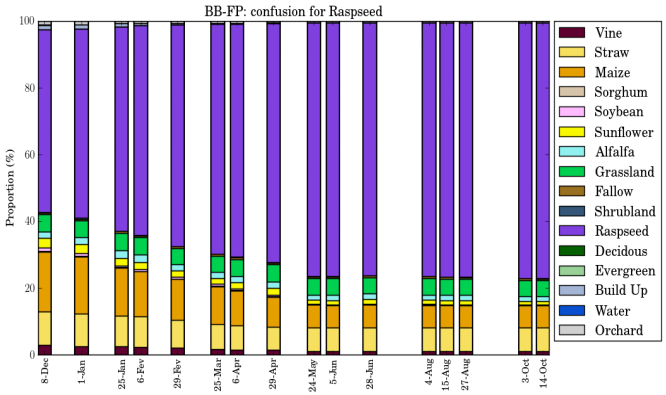
Figure B.43: FN confusion approach for the *Fallow* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.



(a) Sentinel-1

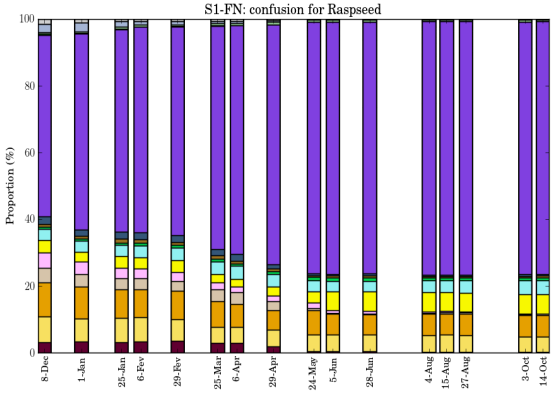


(b) Sentinel-2

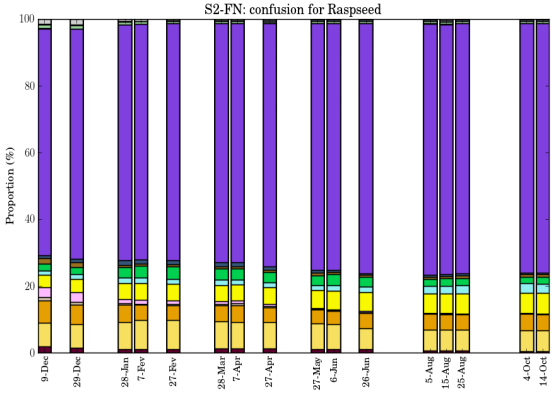


(c) Bayesian Belief Integration

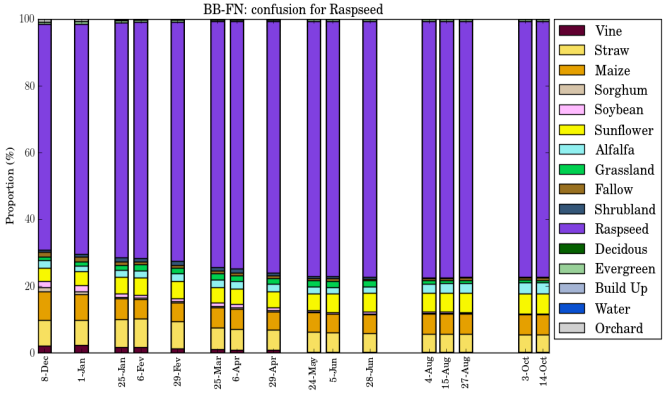
Figure B.44: FP confusion approach for the *Rapeseed* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.



(a) Sentinel-1

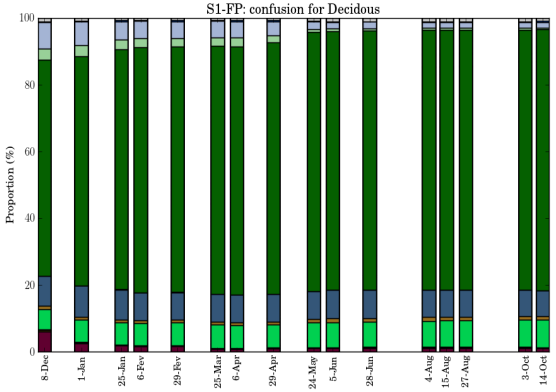


(b) Sentinel-2

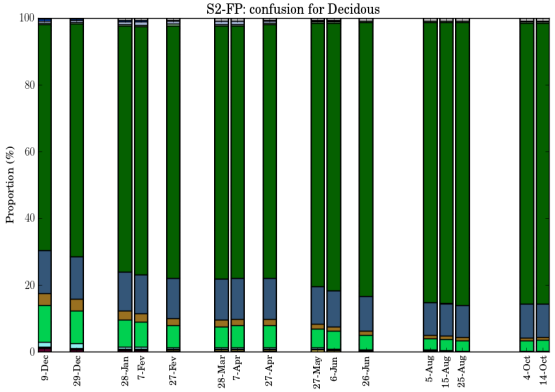


(c) Bayesian Belief Integration

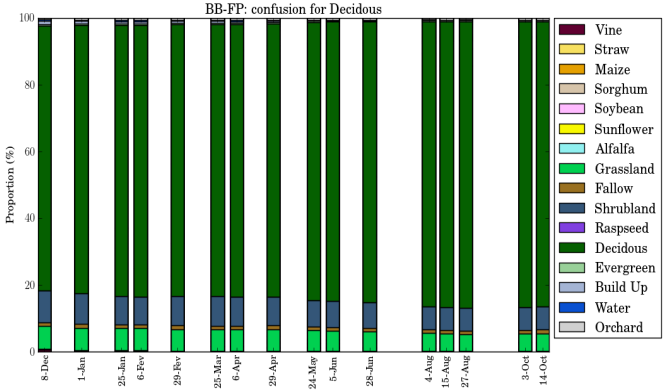
Figure B.45: FN confusion approach for the *Rapeseed* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.



(a) Sentinel-1

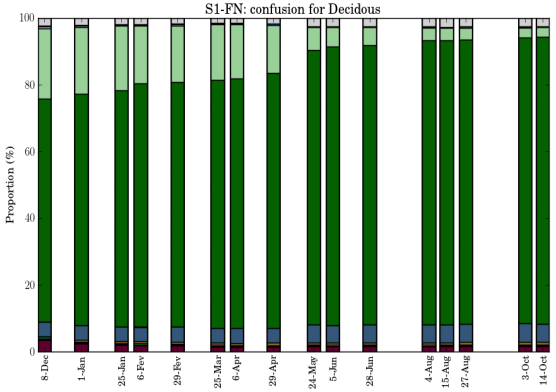


(b) Sentinel-2

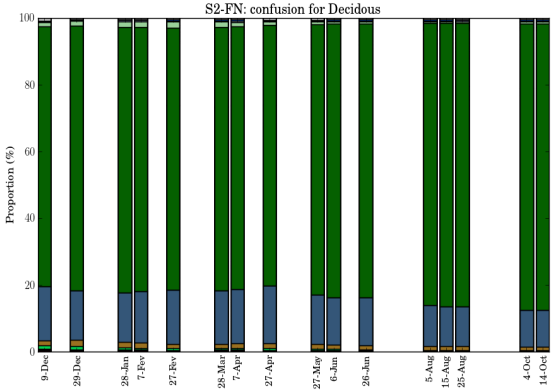


(c) Bayesian Belief Integration

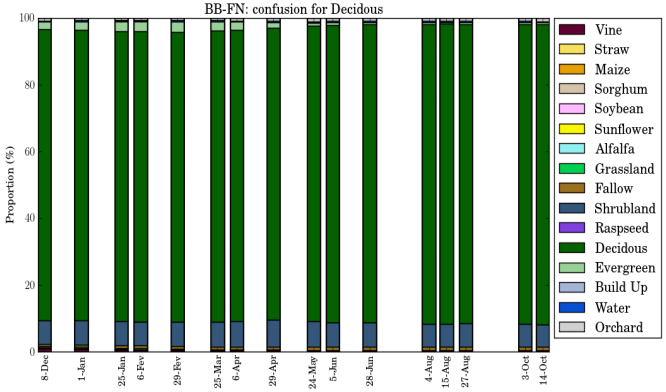
Figure B.46: FP confusion approach for the *Deciduous* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.



(a) Sentinel-1

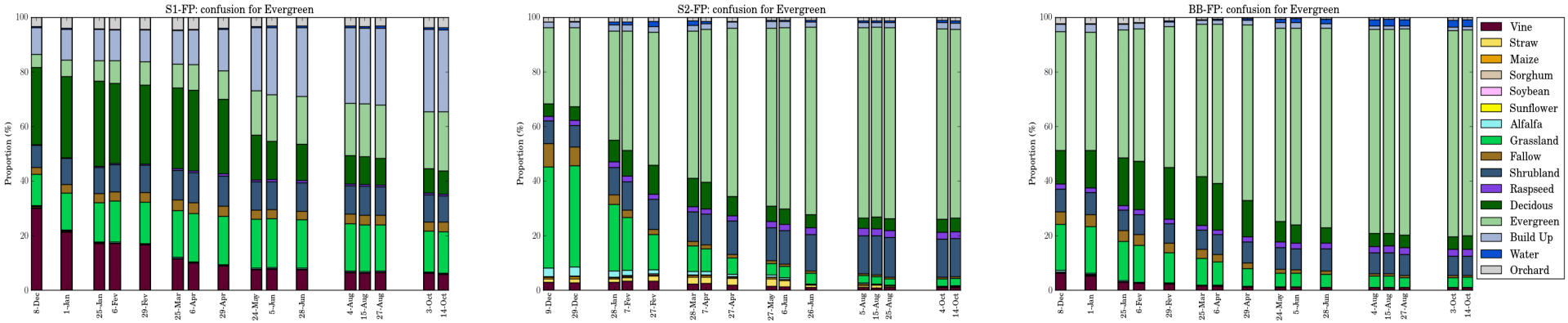


(b) Sentinel-2



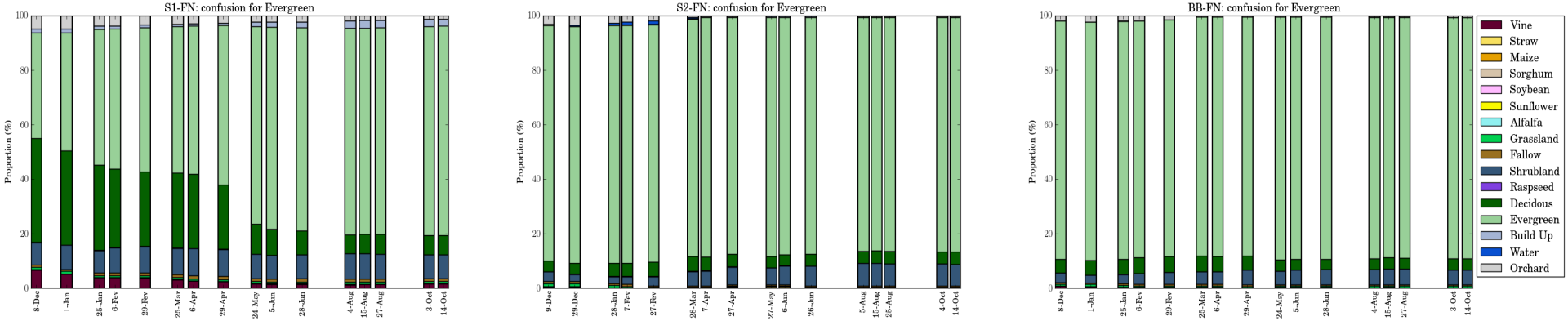
(c) Bayesian Belief Integration

Figure B.47: FN confusion approach for the *Deciduous* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.



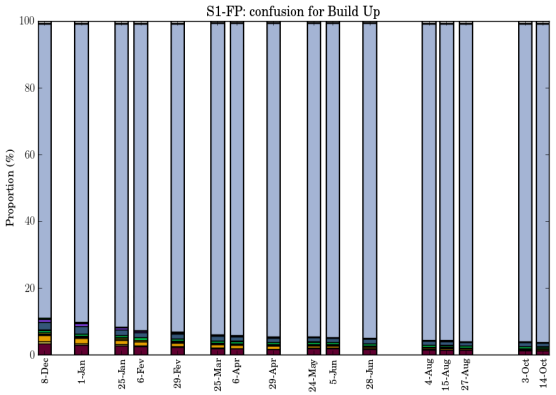
(a) Sentinel-1 (b) Sentinel-2 (c) Bayesian Belief Integration

Figure B.48: FP confusion approach for the *Evergreen* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.

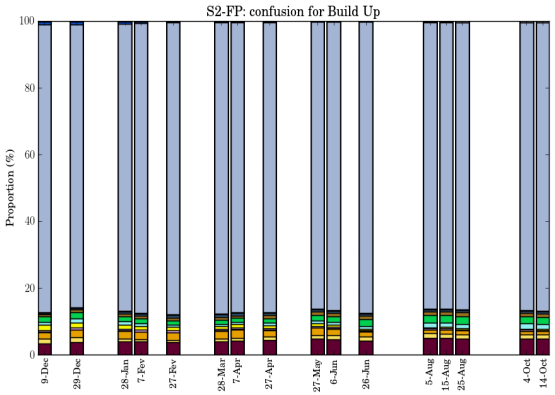


(a) Sentinel-1 (b) Sentinel-2 (c) Bayesian Belief Integration

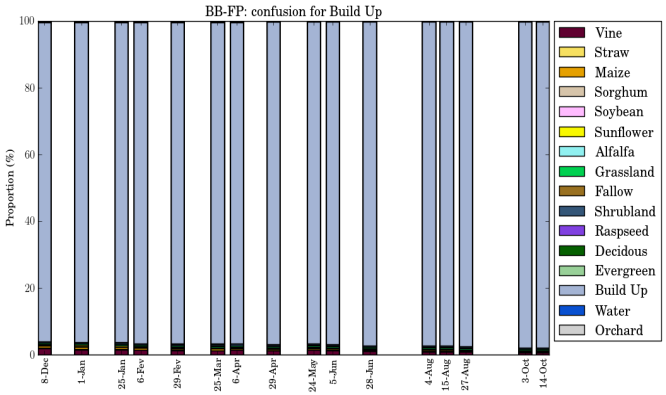
Figure B.49: FN confusion approach for the *Evergreen* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.



(a) Sentinel-1

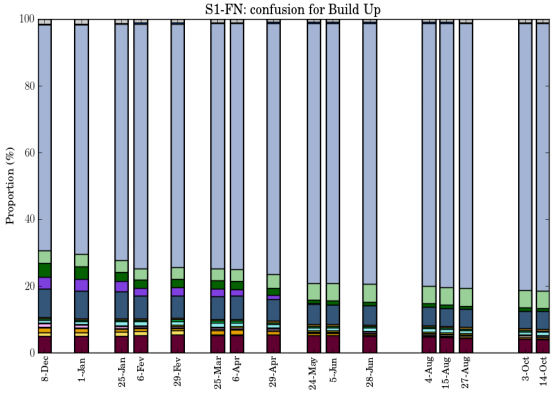


(b) Sentinel-2

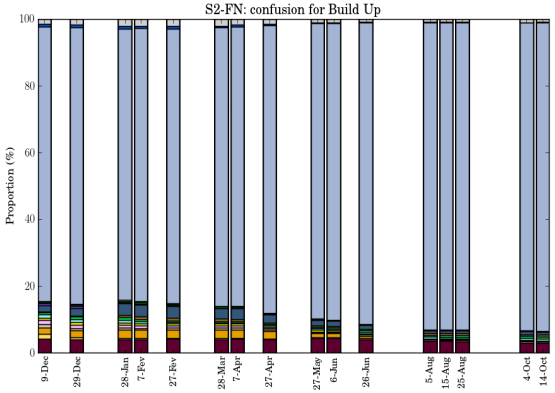


(c) Bayesian Belief Integration

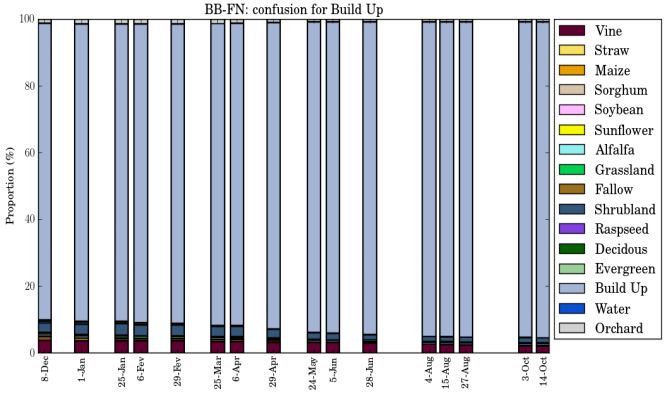
Figure B.50: FP confusion approach for the *Build up* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.



(a) Sentinel-1

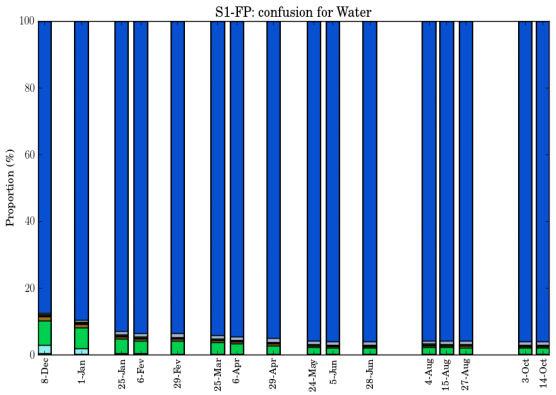


(b) Sentinel-2

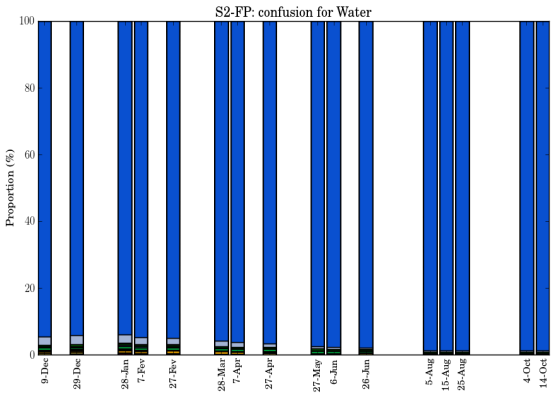


(c) Bayesian Belief Integration

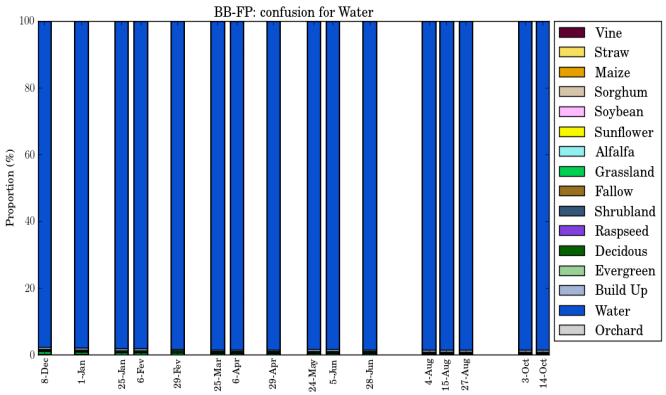
Figure B.51: FN confusion approach for the *Build up* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.



(a) Sentinel-1

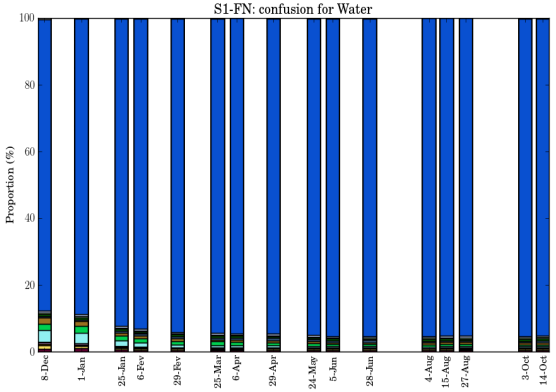


(b) Sentinel-2

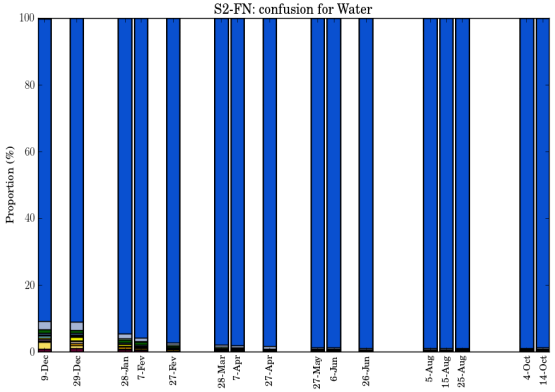


(c) Bayesian Belief Integration

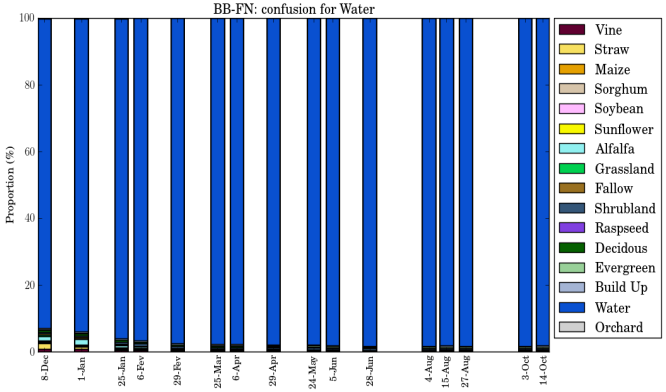
Figure B.52: FP confusion approach for the *Water* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.



(a) Sentinel-1



(b) Sentinel-2



(c) Bayesian Belief Integration

Figure B.53: FN confusion approach for the *Water* class. Temporal evaluation for the (a) S1, (b) S2 and (c) BB strategies.

B.4 Statistical evaluation of the classifications agreements

Classification agreements for the S1, S2 and DS strategies

Class	$S1_{ok}S2_{ok}F_{ok}$	$S1_{ok}S2_{ok}F_{ko}$	$S1_{ok}S2_{ko}F_{ok}$	$S1_{ok}S2_{ko}F_{ko}$	$S1_{ko}S2_{ok}F_{ok}$	$S1_{ko}S2_{ok}F_{ko}$	$S1_{ko} = S2_{ko}F_{ok}$	$S1_{ko} = S2_{ko}F_{ko}$	$S1_{ko} \neq S2_{ko} \mid F_{ok}$	$S1_{ko} \neq S2_{ko} \mid F_{ko}$
Vine	83.28	0.00	4.41	1.75	6.31	0.32	0.00	1.46	0.00	2.48
Straw	86.81	0.00	3.04	0.60	4.95	0.34	0.00	2.44	0.00	1.83
Maize	72.26	0.00	6.89	0.98	9.50	0.22	0.00	5.36	0.00	4.78
Sorghum	42.71	0.00	0.00	16.27	0.03	13.11	0.00	10.25	0.00	17.63
Soybean	59.56	0.00	3.73	8.91	3.27	8.58	0.00	8.62	0.00	7.32
Sunflower	81.57	0.00	2.63	0.46	3.60	1.47	0.00	6.73	0.00	3.55
Alfalfa	44.08	0.00	4.30	13.37	4.10	11.46	0.00	10.29	0.00	12.40
Grassland	39.63	0.00	9.48	1.01	21.55	1.36	0.00	13.56	0.00	13.41
Fallow	35.19	0.00	0.14	16.17	1.67	22.22	0.00	11.15	0.00	13.47
Shrubland	38.60	0.00	3.05	12.55	7.31	12.53	0.00	10.89	0.00	15.06
Raspseed	73.74	0.00	0.83	2.96	0.71	1.15	0.00	14.70	0.00	5.92
Deciduous	78.27	0.00	7.46	0.14	6.86	0.31	0.00	2.90	0.00	4.06
Evergreen	69.60	0.00	0.00	7.43	9.18	7.08	0.00	2.79	0.00	3.92
Build up	76.47	0.00	3.81	0.04	14.95	1.10	0.00	0.73	0.00	2.90
Water	95.04	0.00	0.13	0.00	3.76	0.00	0.00	0.32	0.00	0.75
Orchard	19.08	0.00	0.00	10.55	3.38	24.86	0.00	18.33	0.00	23.80

Table B.1: Statistical evaluation (in %) of the classification agreements for the S1, S2 and DS strategies.

Classification agreements for the S1, S2 and M-DS strategies

Class	$S1_{ok}S2_{ok}F_{ok}$	$S1_{ok}S2_{ok}F_{ko}$	$S1_{ok}S2_{ko}F_{ok}$	$S1_{ok}S2_{ko}F_{ko}$	$S1_{ko}S2_{ok}F_{ok}$	$S1_{ko}S2_{ok}F_{ko}$	$S1_{ko} = S2_{ko}F_{ok}$	$S1_{ko} = S2_{ko}F_{ko}$	$S1_{ko} \neq S2_{ko} \mid F_{ok}$	$S1_{ko} \neq S2_{ko} \mid F_{ko}$
Vine	83.28	0.00	4.91	1.25	5.81	0.81	0.00	1.46	0.00	2.48
Straw	86.81	0.00	3.29	0.34	4.99	0.29	0.00	2.44	0.00	1.83
Maize	72.26	0.00	6.76	1.11	9.28	0.45	0.00	5.36	0.00	4.78
Sorghum	42.71	0.00	0.02	16.26	0.07	13.06	0.00	10.25	0.00	17.63
Soybean	59.56	0.00	3.42	9.22	2.73	9.13	0.00	8.62	0.00	7.32
Sunflower	81.57	0.00	2.42	0.67	4.17	0.89	0.00	6.73	0.00	3.55
Alfalfa	44.08	0.00	3.62	14.06	3.69	11.87	0.00	10.29	0.00	12.40
Grassland	39.63	0.00	9.58	0.92	22.33	0.58	0.00	13.56	0.00	13.41
Fallow	35.19	0.00	1.02	15.29	5.94	17.94	0.00	11.15	0.00	13.47
Shrubland	38.60	0.00	2.96	12.65	6.76	13.08	0.00	10.89	0.00	15.06
Raspseed	73.74	0.00	0.66	3.13	0.84	1.02	0.00	14.70	0.00	5.92
Deciduous	78.27	0.00	7.46	0.13	7.10	0.07	0.00	2.90	0.00	4.06
Evergreen	69.60	0.00	0.10	7.33	12.15	4.11	0.00	2.79	0.00	3.92
Build up	76.47	0.00	3.18	0.67	15.62	0.43	0.00	0.73	0.00	2.90
Water	95.04	0.00	0.12	0.01	3.76	0.00	0.00	0.32	0.00	0.75
Orchard	19.08	0.00	0.00	10.55	6.93	21.31	0.00	18.33	0.00	23.80

Table B.2: Statistical evaluation (in %) of the classification agreements for the S1, S2 and M-DS strategies.

Classification agreements for the S1, S2 and MC strategies

Class	$S1_{ok}S2_{ok}F_{ok}$	$S1_{ok}S2_{ok}F_{ko}$	$S1_{ok}S2_{ko}F_{ok}$	$S1_{ok}S2_{ko}F_{ko}$	$S1_{ko}S2_{ok}F_{ok}$	$S1_{ko}S2_{ok}F_{ko}$	$S1_{ko} = S2_{ko}F_{ok}$	$S1_{ko} = S2_{ko}F_{ko}$	$S1_{ko} \neq S2_{ko} \mid F_{ok}$	$S1_{ko} \neq S2_{ko} \mid F_{ko}$
Vine	83.28	0.00	3.97	2.19	5.92	0.70	0.00	1.46	0.00	2.48
Straw	86.81	0.00	2.92	0.72	4.74	0.54	0.00	2.44	0.00	1.83
Maize	72.26	0.00	4.80	3.08	8.24	1.48	0.00	5.36	0.00	4.78
Sorghum	42.71	0.00	10.25	6.02	7.85	5.28	0.00	10.25	0.00	17.63
Soybean	59.56	0.00	8.12	4.52	8.38	3.47	0.00	8.62	0.00	7.32
Sunflower	81.57	0.00	2.46	0.63	3.72	1.34	0.00	6.73	0.00	3.55
Alfalfa	44.08	0.00	8.64	9.04	10.52	5.04	0.00	10.29	0.00	12.40
Grassland	39.63	0.00	5.84	4.65	16.54	6.37	0.00	13.56	0.00	13.41
Fallow	35.19	0.00	9.18	7.13	18.33	5.55	0.00	11.15	0.00	13.47
Shrubland	38.60	0.00	4.73	10.87	16.19	3.65	0.00	10.89	0.00	15.06
Raspseed	73.74	0.00	3.42	0.37	1.23	0.63	0.00	14.70	0.00	5.92
Deciduous	78.27	0.00	5.07	2.53	6.67	0.50	0.00	2.90	0.00	4.06
Evergreen	69.60	0.00	2.83	4.61	15.43	0.84	0.00	2.79	0.00	3.92
Build up	76.47	0.00	3.02	0.84	14.79	1.26	0.00	0.73	0.00	2.90
Water	95.04	0.00	0.11	0.02	3.46	0.30	0.00	0.32	0.00	0.75
Orchard	19.08	0.00	4.49	6.06	20.37	7.87	0.00	18.33	0.00	23.80

Table B.3: Statistical evaluation (in %) of the classification agreements for the S1, S2 and MC strategies.

Classification agreements for the S1, S2 and MR strategies

Class	$S1_{ok}S2_{ok}F_{ok}$	$S1_{ok}S2_{ok}F_{ko}$	$S1_{ok}S2_{ko}F_{ok}$	$S1_{ok}S2_{ko}F_{ko}$	$S1_{ko}S2_{ok}F_{ok}$	$S1_{ko}S2_{ok}F_{ko}$	$S1_{ko} = S2_{ko}F_{ok}$	$S1_{ko} = S2_{ko}F_{ko}$	$S1_{ko} \neq S2_{ko} \mid F_{ok}$	$S1_{ko} \neq S2_{ko} \mid F_{ko}$
Vine	83.28	0.00	4.64	1.52	6.17	0.45	0.00	1.46	0.29	2.19
Straw	86.81	0.00	3.17	0.46	4.72	0.57	0.00	2.44	0.12	1.71
Maize	72.26	0.00	5.53	2.34	8.76	0.97	0.00	5.36	0.50	4.28
Sorghum	42.71	0.00	11.17	5.10	7.55	5.58	0.00	10.25	1.22	16.42
Soybean	59.56	0.00	8.81	3.83	8.05	3.80	0.00	8.62	0.46	6.86
Sunflower	81.57	0.00	2.38	0.71	3.61	1.45	0.00	6.73	0.08	3.47
Alfalfa	44.08	0.00	10.54	7.13	11.51	4.05	0.00	10.29	1.24	11.16
Grassland	39.63	0.00	5.44	5.05	17.79	5.12	0.00	13.56	0.74	12.67
Fallow	35.19	0.00	8.99	7.32	19.77	4.12	0.00	11.15	0.96	12.51
Shrubland	38.60	0.00	9.40	6.21	15.80	4.04	0.00	10.89	1.00	14.06
Raspseed	73.74	0.00	3.51	0.28	1.09	0.77	0.00	14.70	0.03	5.89
Deciduous	78.27	0.00	6.01	1.58	5.86	1.31	0.00	2.90	0.19	3.87
Evergreen	69.60	0.00	3.51	3.92	15.12	1.14	0.00	2.79	0.18	3.74
Build up	76.47	0.00	3.27	0.58	14.97	1.08	0.00	0.73	0.34	2.57
Water	95.04	0.00	0.11	0.02	3.36	0.40	0.00	0.32	0.01	0.74
Orchard	19.08	0.00	5.19	5.37	23.15	5.09	0.00	18.33	2.24	21.55

Table B.4: Statistical evaluation (in %) of the classification agreements for the S1, S2 and MC strategies.

B.5 Visual evaluation of the classifications agreements

Dempster-Shafer agreement map

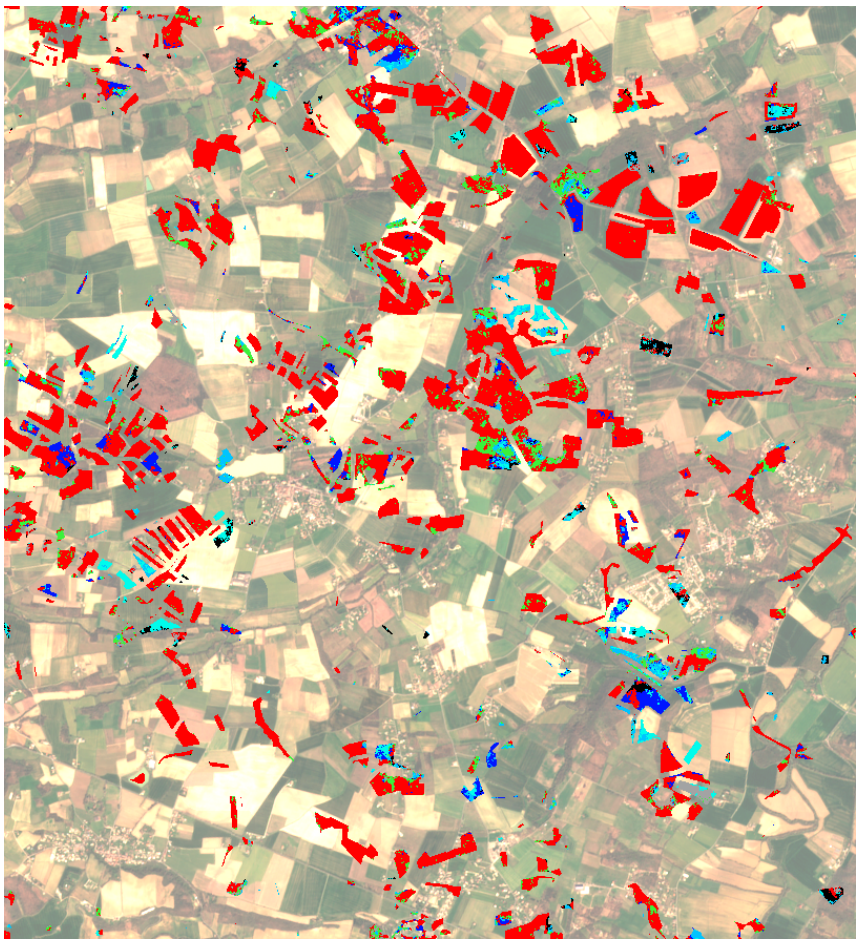


Figure B.54: Classifications agreement map for the DS fusion approach (14th October).

Modified Dempster-Shafer agreement map

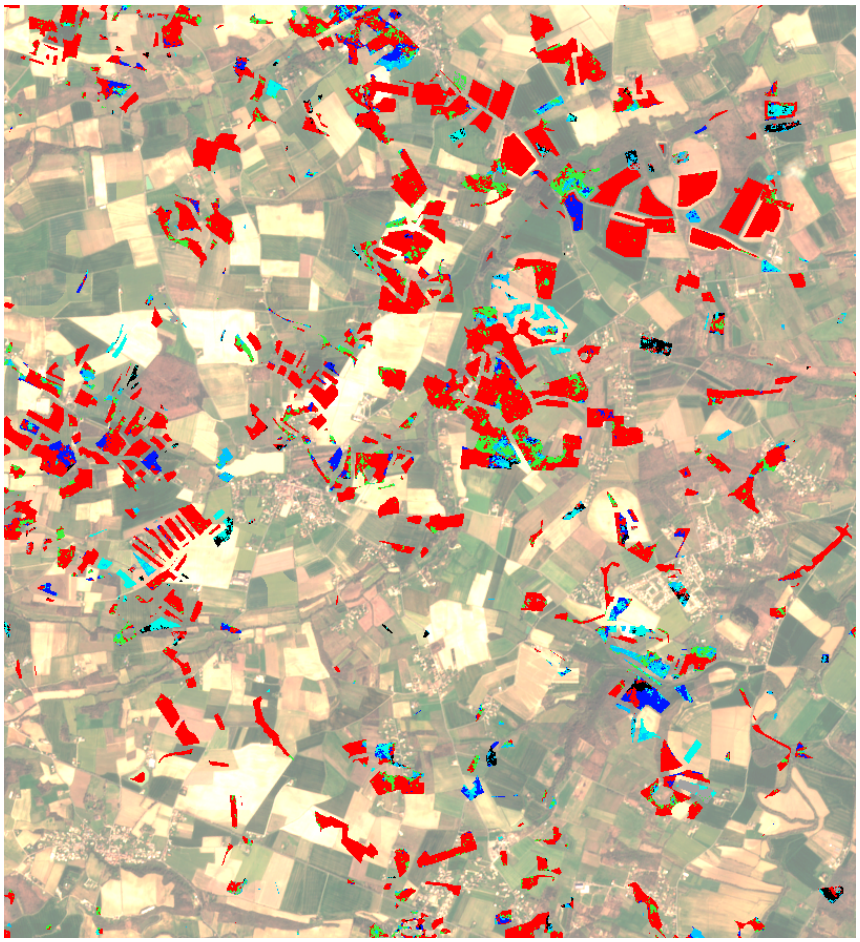


Figure B.55: Classifications agreement map for the M-DS fusion approach (14th October).

Maximum Confidence agreement map

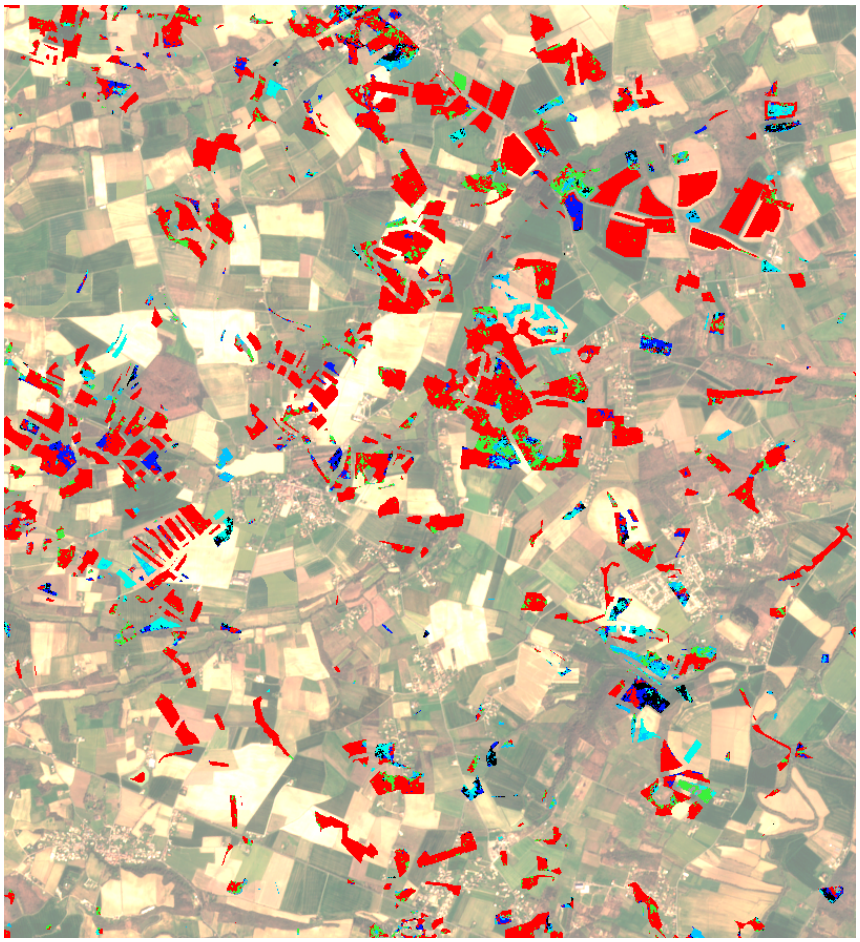


Figure B.56: Classifications agreement map for the MC fusion approach (14th October).

Median Rule agreement map

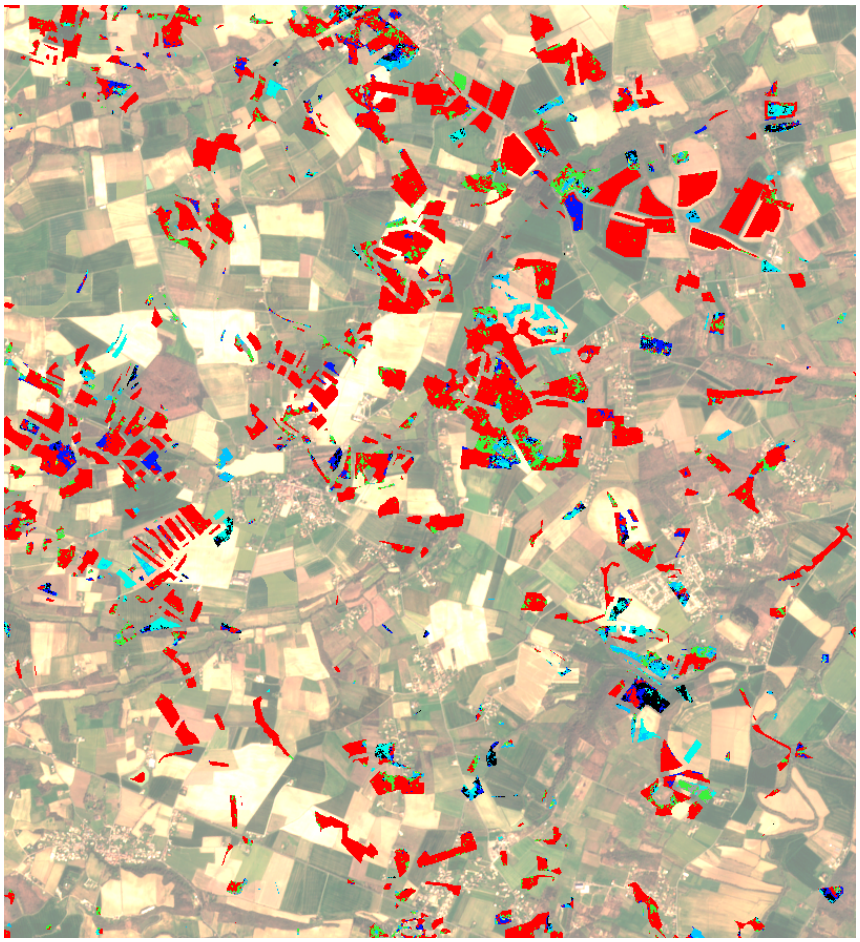


Figure B.57: Classifications agreement map for the MR fusion approach (14th October).

B.6 Fusion strategies summary

B.6.1 Operation principle

Fusion method	Basic principle
Dempster-Shafer	For a given pixel, given the confusion matrices and the predicted labels from radar and optical classifiers, a parameter called *belief* is calculated per each predicted label by means of the Dempster-Shafer theory of evidence, being a generalization of the Bayesian reasoning. This parameter expresses the level of uncertainty of belonging to a given class. Thus, the rule assigns the fused label to the predicted label reaching the maximum belief.
Modified Dempster-Shafer	It applies the same rules than the aforementioned Dempster-Shafer method but a new version of the belief measure is propose. The *belief* parameter is computed as DS fusion does but the values are weighted. This weighting is carried out by means of the probability corresponded to each predicted class.
Bayesian Belief Integration	For a given pixel, given the probability vectors from radar and optical classifiers, a parameter called *belief* is calculated. This metric consist in an array expressing the aggregated belief per class. This parameter expresses the level of uncertainty of belonging to a given class. Thus, the rule assigns the predicted label to the class with maximum belief.
Maximum Confidence	For a given pixel, given the probability vectors from radar and optical classifiers, the fusion criterion assigns the fused label to the class with highest probability. The most straightforward method.
Median Rule	For a given pixel, a combined probability vector is obtained by averaging the probability vectors from radar and optical classifiers. Thus, the rule assigns the fused label to the class with highest value after the averaging.

B.6.2 Advantages

Fusion method	Advantages

Dempster-Shafer	– Apply a weight to each classifier depending on its accuracy – No need of probability vectors – Widely adopted evidence theory – Low computational burden – Best precision and f-score results for Sorghum class
Modified Dempster-Shafer	– Apply a weight to each classifier depending on its accuracy – It exploits the highest probability value for a given prediction – Widely adopted evidence theory – Low computational burden – Best recall results for Straw, Sunflower, Grassland, Deciduous, Build up and Water classes (eliminates a large amount of false negatives)
Bayesian Belief Integration	– It achieves the best results – No class label dependency – Takes full profit from the RF probabilities – Works better with uncertain probability vectors (product shows higher differences between values than sum) – Strong and widespread inference theorem behind – Low computational burden

Maximum Confidence	– It exploits the highest RF probability value – Low computational burden (The most straightforward method)
Median Rule	– It achieves excellent results – No class label dependency – Takes full profit from the RF probabilities – Does not depend on the composition of the probability vectors (not affected by zero values) – Low computational burden

B.6.3 Drawbacks

Fusion method	Drawbacks
Dempster-Shafer	– It performs the decisions based on the predicted labels – The *masses of Belief* are established based only on one metric (the precision for the best case) – If both input labels are wrong, the output label is wrong
Modified Dempster-Shafer	– It performs the decisions based on the predicted labels – The masses of Belief are established based only on one metric (the precision for the best case) – If both input labels are wrong, the output label is wrong – It only exploits the maximum probability for each classifier – Given a probability vector, if two classes have the same maximum probability the choice is random

Bayesian Belief Integration

- If the classifiers predict with a higher level of confidence, the probability vectors can be similar to a delta function. Thus if it happens for both classifiers but with different classes the result is a product of 0's (product between delta functions for different X values)
- The probabilistic approach is quite straightforward
- Probabilistic combiners present poor precision results for some classes such as Orchard, Fallow or Sorghum where the single classifiers present unbalanced results
- All the classifiers have the same weight in the decision process

Maximum Confidence

- Given a probability vector, if two classes have the same maximum probability the choice is random
- It only exploits the maximum probability value from the RF probability vector
- Probabilistic combiners present poor precision results for some classes such as Orchard, Fallow or Sorghum
- All the classifiers have the same weight in the decision process

Median Rule	– All the classifiers have the same weight in the decision process – Sum operand shows less discrimination than product – Probabilistic combiners present poor precision results for some classes such as Orchard, Fallow or Sorghum where the single classifiers present unbalanced results